Development of Cost-effective Single Reference Methods for an Accurate Description of Dynamic Correlation Energy

Dissertation
zur
Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. Nat.)

vorgelegt der
Mathematisch-naturwissenschaftlichen Fakultät
der
Universität Zürich

von
Loïc Roch
aus
Lausanne, VD

Promotionskomitee
Prof. Dr. Kim K. Baldridge (Vorsitz)
Prof. Dr. Karl-Heinz Ernst
Prof. Dr. Jay S. Siegel
Prof. Dr. Peter R. Taylor
Prof. Dr. Daniel Tunega

September 22, 2016

Abstract

Computational chemistry plays a major role in providing new conceptual approaches for a better understanding of atoms, molecules and materials systems in general. Typically, the role of theory has been considerable in modern applications for prediction of structure as well as a wide range of chemical and physical properties, and has had remarkably accelerated progress in many areas, ranging from small molecule reaction processes to industrial-scale processes. This has in turn lead to major developments as well as discoveries in chemistry. Despite this, there are still important gaps in what is presently available in term of quantum chemical methods that need to be filled to tackle other relevant problems. Notably, it is still necessary to (i) constantly re-think effort in developing new methods to reduce computational cost without scarifying high accuracy, (ii) collect experimental data that can exemplify appropriately, and, thereby also, validate new theoretical methods, and (iii) design high performance computer infrastructure able to carry out calculations at high levels of accuracy. The work of this thesis encapsulates these three challenges, with particular emphasis on the development and implementation of several cost-effective single reference methods for accurate description of electron correlation. Spanning material sciences to biology, the effects of electron correlation are ubiquitous: packing of molecules into solids, self-assembly, molecular recognition, three-dimensional structures of proteins, interaction of substrates with surfaces, homogeneous and heterogeneous catalysis, to name but a few. With new techniques such as those presented in this work, one can make accurate predictions on aspects of structure and properties of large systems, and thereby provide an important supplement to experimental chemistry with added detail and even prediction of yet unknown chemical outcomes. The work is exemplified in important new and current chemical challenges.

Zusammenfassung

Computergestützte Chemie spielt eine wichtige Rolle, um neue Ansätze für ein besseres Verständnis der Atome, Moleküle und Materialsysteme im Allgemeinen zu entwickeln. Die Theorie, die für moderne Anwendungen einen hohen Stellenwert in der Vorhersage von Molekularstrukturen sowie von chemischen und physikalischen Eigenschaften besitzt, hat in den letzten Jahren in vielen Bereichen, von simplen chemischen Reaktionen bis hin zu komplexen Industrieprozessen, eine rasante Entwicklung erfahren. Dies führte zu bedeutenden Fortschritten sowie neuen Entdeckungen in der Chemie. Trotzdem bestehen immer noch Lücken in den Methoden der Quantenchemie die geschlossen werden müssen. Vor allem ist es notwendig (i) neue Methoden zu entwickeln, um die Berechnungszeit zu minimieren und damit die entstehenden Kosten zu senken ohne jedoch Genauigkeit einzubüssen, (ii) experimentelle Daten zu sammeln, welche neu entwickelte Theorien unterstützen und validieren können, (iii) neue Hochleistungscomputerinfrastruktur zu entwickeln, die in der Lage ist die benötigten Berechnungen mit einer hohen Genauigkeiten durchzuführen. Die vorliegende Dissertation beinhaltet diese drei genannten Problemstellungen und legt dabei einen besonderen Schwerpunkt auf die Entwicklung und Implementierung von mehreren kosteneffizienten Single-reference Methoden zur präzisen Bestimmung der Elektronenkorrelation. Der Einfluss der Elektronenkorrelation ist von der Materialwissenschaft bis hin zur Biologie allgegenwärtig: Anordnung von Molekülen in Festkörpern, Selbstassemblierung, molekulare Erkennung, drei-dimensionale Struktur von Proteinen, Wechselwirkungen zwischen Substraten und Oberflächen, homogene und heterogene Katalyse sowie eine grosse Anzahl weiterer Aspekte. Mit Verfahren wie sie in diesem Bericht vorgestellt werden, können genaue Vorhersagen über die Struktur und die Eigenschaften von grossen Systemen gemacht werden und sie stellen damit eine wichtige Ergänzungen zur experimentellen Chemie dar. Darüber hinaus können Prognosen für bis dato unbekannte chemische Ergebnisse gemacht werden. Die in dieser Arbeit vorgestellten Methoden werden dabei auf ausgewählte, aktuelle chemische Problemstellungen angewendet.

Contents

« You can't stay in your corner of the Forest waiting for others to come to you. You have to go to them sometimes.»

A. A. Milne

Chapter 1

Introduction

Computational chemistry is a sub-field of chemistry which brings together theory and computer science. The primary focus of computational chemistry is on understanding the fundamental properties of atoms, molecules, and matter, using theories arising from quantum mechanics. Typically, theory finds considerable modern applications in predicting various properties such as, *e.g.*, NMR chemical shifts, electronic spectra, electrochemical properties, or thermodynamics. Notably the use of computational chemistry incredibly increased as computational hardware and software became more powerful and cheaper.[1] This astonishing pace profoundly influenced conceptual approaches in understanding chemistry. As a consequence, the synergy between theoretical predictions and experimental observations has remarkably accelerated progress in many areas ranging from atoms and molecules to industrial-scale processes, leading to major discoveries.[2]

The work of this thesis involved the development, implementation, and application of several cost-effective single reference methods for accurate calculations of molecular energies and properties of real systems. Despite the myriad of methodologies that is currently available, there are still important gaps in the capabilities in quantum chemical methods that need to be filled in order to tackle relevant problems of interest today. In particular, aiming at reaching the highest level of prediction requires effort in developing new quantum chemical methods, and in improving or supplementing existing methods, appropriate and reliable for specific needs. In addition, having experimental applications that can exemplify appropriately, and, thus, validate such new theoretical treatments, is of utmost importance. Once validated, such methods can fill in areas, which remained uncleared in the experimental knowledge for prediction of fundamental properties.

Although the performance of the machinery used for computations – *i.e.* High Performance Computers (HPC) infrastructures – has considerably improved over the past few years, recently reaching the petaflops scale, quantum chemical methods scale exponentially with the system size limiting their scope of applications to fairly small systems (*ca.* 800 basis functions). This limitation leads to efforts in designing highly available HPC able to carry out calculations at adequate levels of theory for predictable phenomenon. Furthermore, this work aims at developing new methods capable of handling large systems, *i.e.*, above ca. 800 basis functions. This means constantly re-thinking the development of new methods to reduce the computational cost without sacrificing high accuracy.

As such, the work of this thesis encapsulates three challenges, which can be summarized as:

(A) Quantum mechanical challenges: formulation of new and more accurate methods that enable a high level of predictability.

(B) Hardware challenges: design of hardware systems that are able to carry out the high end computation laid out by the heavy methods and large system size.

(C) Chemical challenges: ability to understand relevant chemical applications at a level, which enables high predictability of specific phenomenon and ensures appropriate interpretation of the calculated data.

For a deep understanding of the different components involved in the sub-field of computational chemistry, the remainder of this Ph.D. thesis covers these challenges.

A. Quantum mechanical challenge

For the last 50 years, electron correlation has been at the heart of atomic, molecular, and solid-state theory.[3] Although the correlation energy, also known as non-covalent interactions, is typically *ca.* 1% of

the total energy, a correct description of electron correlation is of utmost importance for the prediction of chemical and physical properties in most applications. Despite the significant achievements made since the early days of computational chemistry in the prediction of material properties, an accurate description of weak interactions in pertinent systems is still very difficult. Traditionally, a reasonable description of correlation effects is achieved by application of Møller-Plesset second order perturbation theory[4] (MP2) or coupled-cluster methods[5–7] (CC). In general, in either case, computational costs limit their scope of application to small systems (*ca.* 800 basis functions), warranting efforts to design methods capable of reducing the computational cost. Directed at this point, the current work explores a new family of double-hybrids density functional theory[8;9] (DFT) that handles the electron correlation and enables enhanced accuracy and scalability.

Double-hybrid density DFT appears as a promising technique to account for the incorrect long range asymptotic behavior of standard functionals: this technique mixes an exact exchange term with the approximate DFT exchange functional as simple hybrids, but adds a perturbational correlation term to the approximate DFT correlation in the basis of the Kohn-Sham orbitals.[10;11]

As a first approach to density functional development, the dispersion-corrected spin-component-scaled double-hybrids[12;13] (DSD) DFTs were implemented into the General Atomic Molecular Electronic Structure Systems[14] (GAMESS) and were intensively studied. This effort includes the development of validation test sets categorized in accordance with specific stabilizing contributions. In this regard, 66 chemical systems with correlation ranging from 18.6 kcal/mol to 0.02 kcal/mol were selected and grouped into distinct classes.

The implementation of DSD-DFs lead to the development of a new type of double-hybrid DFs. Aimed at minimizing the number of empirical parameters, the spin-component-scaled (SCS) DFs, although still semi-empirical in character, possesse only two fitted parameters (for reference, there are five fitted parameters in the DSD-DFs). The performance of such scheme is shown to provide results of similar accuracy to the DSD-DF that out-performs the MP2 method in most cases.

On the way towards improving the performance and reducing the computational cost of the double-hybrid functionals, the range-separated MP2 perturbation theory developed by M. Head-Gordon *et al.*[15] has also been implemented into GAMESS. After validating the implementation of the range-separated MP2, the latter perturbative treatment of the correlation energy was extended to the approximate correlation functional, leading to the double-hybrid $erfc$ functionals, further referred to as DH($erfc$)-DFs.

Further, for cost-effectiveness, the newly implemented methodologies have also been RI-enabled.[16] In this way, the computationally expensive four index two-electron integrals are solved within the resolution-of-identity approximation. An exhaustive auxiliary basis set analysis was also performed as a part of this effort, in order to assess the minimal requirements.

All of the above-mentioned quantum chemical challenges are covered in Chapter 2, which starts with an overview of the available quantum methods and a general note on their corresponding performance (section I). Then the various theories used for developing new methodologies are presented in sections II and III. Section IV describes and motivates the seven new data sets developed in this work. An exhaustive basis set convergence study and discussion of the results follows in section V. Then, the performance and implementation of the DSD-DFTs is discussed in section VII. The details of the implementation and optimization procedure of the new SCS-DFTs are summarized in section VIII.

The performance of the resolution-of-identity approach with the double-hybrid frame is discussed in section IX. Finally, the implementation of the range-separated MP2 is detailed in section X and the merge with the hybrid DFs is explained in section XI.

B. Hardware challenge

During the last decade, high-performance computing (HPC) has continued expending at an astonishing exponential pace[1;17–19] to become an important resource for scientists world-wide for the purpose of understanding problems with computer simulations of real-world applications, spanning climate dynamics, engineering, astrophysics, nanotechnology, chemistry, and biophysics, to name but a few.[20;21] Lately, supercomputers, in particular on the petaflop scale, have become even more prevalent in academia. For example, to date, eleven of the 95 elite petaflop machines are installed at universities, not all of which are associated with deep-pocketed supercomputing centers. As a matter of fact, it is becoming increasingly common for at least large universities to acquire their own peta-machines for in-house researchers, rather than being dependent on the charity of national labs to share such resources.[22] The number of challenges for system software and scientific applications with respect to reliability, availability and serviceability has considerably increased with the resources growth.

In this aspect, the race for scientific discovery by running applications on the fastest machines available for a significant amount of time (*i.e.* weeks and months), while demanding high throughput without interruption, has forced a re-design of high-performance computing (HPC) infrastructures. Consequently, the search for fault-tolerant highly available HPC for large parallel quantum chemistry calculations is of utmost importance. However defining a level of redundancy is strategic when planning a new data center as it directly impacts the entire design of the building as well as the construction and operational costs. It also affects how to integrate future extension plans into the design.[23] The downside of redundancy is that extra resources are required and there is an additional overhead on communication and synchronization.[24] With the notion of two-level redundancy, availability of HPC is considerably increased. In general, large-scale HPC systems[22] may be partitioned, separate interconnected networks may exist to minimize interference, user data and authentication services may be mirrored, with the purpose of maximizing the overall reliability.

Chapter 4 details the design, development and mass deployment of a highly available HPC for large quantum chemistry calculations. In a first time the general layout of the HPC infrastructure is shown and the corresponding components are detailed (section III and IV). The notion of redundancy at the power supply level is briefly introduced in section V. The importance of a highly available storage system and of a scale-out storage is described in section VI. Also, the performance of such scale-out storage is detailed in section XII. Section VII is an exhaustive description of the networks connectivity within the cluster and to the outside world. The critical parts harnessed by the resource manager and its configuration are outlined in section VIII. The automated tools used for mass deployment of the configurations are listed in section IX. Finally, the importance of health checks is highlighted in section X.

C. Chemical challenge

Advances in computing have facilitated the synergy between theoretical predictions and experimental observations, which has lead to major discoveries. Indeed, on the one side, experimental facts without a theoretical interpretation often do not provide enough of a detailed understanding, and on the other

side theory without a comparison with experiment can lead to unrealistic dreams.[1]

Chapter 3 illustrates a synergistic study combining experiment and theory. Specifically, this investigation relates the properties of a new polymorph system, pentaindenocorannulene and its ability to form complexes and aggregate with C_{60}. The interactions between aromatic moieties such as these are of concave-convex π-π interactions nature, and as such fit well into the general theme of this thesis. In the first section the structural parameters, such as cone angles and interacting distances are presented and compared to experimental results. It is shown that computational chemistry is able to elucidate the exact orientation of C_{60} within the heavily disordered aggregate. In addition, theory facilitated the assignment of the indistinct fourth anionic state of pentaindenocorannulene (section IIIC.), as well as the structural issues associated with the experimental NMR (section IIIB.). The competition between aggregation and complexation is exhaustively studied by means of NMR prediction and TD-DFT calculations, in sections IIIB. and IIID., respectively.

Chapter 2

Theory and methods development

I Overview of the available quantum chemical methods

The present section is an overview of the available quantum chemical methods. It starts with a summary of the main results obtained from the derivation of the Hartree-Fock equation, which is an important starting point for the development of more accurate approximations including correlation effects, namely the *post* Hartree-Fock methods. Møller-Plesset second order perturbation theory (MP2), and MP2-like methods are then presented and derived in detail. Before presenting the Density Functional Theory (DFT), a short overview of the coupled-cluster (CC) *ansatz* and the CC equations are presented.

All the theory presented in this section is inspired from references [3; 25–31].

A. Hartree-Fock approximation

The Hartree-Fock (HF) approximation[32] is central to chemistry and is at the origin of more accurate approximations, which account for electron correlation. The HF theory is interested in finding a set of spin orbitals $\{\chi_a\}$ such that the single (Slater) determinant formed from these spin orbitals

$$|\Psi_0\rangle = |\chi_1\chi_2\cdots\chi_i\chi_j\cdots\chi_N\rangle \tag{2.1}$$

is the best possible approximation to the ground state of the N-electron system described by the non-relativistic Born-Oppenheimer electronic Hamiltonian $\hat{H}_{el}$.

$$\hat{H}_{el} = -\frac{1}{2}\sum_{i=1}^{N}\nabla_i^2 - \sum_{i=1}^{N}\sum_{A=1}^{M}\frac{Z_A}{|r_i - R_A|} + \sum_{\substack{i=1\\j>i}}^{N}\frac{1}{|r_i - r_j|} + \sum_{\substack{A=1\\B>A}}^{M}\frac{Z_A Z_B}{|R_A - R_B|} \tag{2.2}$$

where Z_A denotes the charge of nucleus A, ∇_i the gradient operator for particle i, such as

$$\nabla_i = \left(\frac{\partial}{\partial x_i}, \frac{\partial}{\partial y_i}, \frac{\partial}{\partial z_i}\right) \tag{2.3}$$

and

$$\nabla_i^2 = \left(\frac{\partial^2}{\partial x_i^2} + \frac{\partial^2}{\partial y_i^2} + \frac{\partial^2}{\partial z_i^2}\right) \tag{2.4}$$

Notice that in eq. 2.2 the separation of electrons and nuclei is *not* symmetric. The electronic hamiltonian $\hat{H}_{el}$ depends parametrically on the nuclear positions, and, thus, appears in the $\hat{H}_{el}$, whereas derivatives with respect to these coordinates do not. As a consequence, the Schrödinger electronic problem

$$\hat{H}_{el}(r, R)\,\psi_{el}(r, R) = E_{el}(R)\,\psi_{el}(r, R) \tag{2.5}$$

is solved for a set of nuclear coordinates R which are momentarily clamped to fixed positions in space. According to the variational principle, the best $\{\chi_a\}$ is the one minimizing the electronic energy E_0.

$$\begin{aligned} E_0 = \left\langle \Psi_0|\hat{H}_{el}|\Psi_0 \right\rangle &= \sum_i \left\langle \chi_i \left| \hat{h} \right| \chi_i \right\rangle + \frac{1}{2}\sum_{\substack{i\\j\neq i}} \left(\langle \chi_i\chi_j \left|\hat{g}\right| \chi_i\chi_j\rangle - \langle \chi_i\chi_j \left|\hat{g}\right| \chi_j\chi_i\rangle\right) \\ &= \sum_i \left\langle \chi_i \left| \hat{h} \right| \chi_i \right\rangle + \frac{1}{2}\sum_i \left\langle \chi_i \left| \hat{J} - \hat{K} \right| \chi_i \right\rangle \\ &= \sum_i \underline{\left\langle i \left| \hat{h} \right| i \right\rangle} + \frac{1}{2}\sum_{\substack{i\\j\neq i}} \underline{\langle ij|ij\rangle} - \underline{\langle ij|ji\rangle} \end{aligned} \tag{2.6}$$

Where the expression $\hat{h} + \hat{J} - \hat{K}$ is the so-called Fock operator:

$$\hat{F} = \hat{h} + \hat{J} - \hat{K} \tag{2.7}$$

It is worth mentioning that $\hat{h}$ depends on the coordinates of a single electron (*i.e.* the first two terms in eq. 2.2), and that both the Coulomb and the exchange operators ($\hat{J}$ and $\hat{K}$, respectively) arise from the two-electron part of eq. 2.2. Schematically, E_0 is described by Figure 2.1.

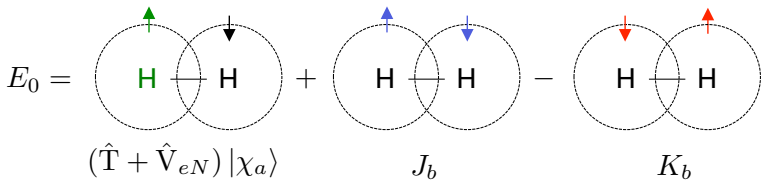


Figure 2.1: Schematic representation of the HF energy. The colors refer to the colors in eq. 2.6

B. Møller-Plesset second-order perturbation theory

In the Møller-Plesset second-order perturbation theory[4] (MP2) approach, the Hamiltonian $\hat{H}$ is divided into an unperturbed part $\hat{H}_0$, the HF Hamiltonian (eq. 2.2), and a perturbation $\hat{V}$ which is defined as:

$$\hat{V} = \sum_{\substack{i \\ j>i}} \frac{1}{|r_i - r_j|} - \sum_i v^{HF}(i) \tag{2.8}$$

leading to

$$\hat{H} = \hat{H}_0 + \lambda \hat{V} \tag{2.9}$$

where λ is an ordering parameter, and $\sum_i v^{HF}(i)$ is the HF potential, an effective one-electron potential operator.

The total energy E, and eigenfunctions $|\psi\rangle$ (see eq. 2.5) are then expanded in a similar way and it is assumed that the zeroth order term $|\psi_{(0)}\rangle$ is an eigenfunction of $\hat{H}_0$ with $E_{(0)}$ as eigenvalue.

$$\begin{aligned} E &= E_{(0)} + \lambda E_{(1)} + \lambda^2 E_{(2)} + \cdots \\ \psi &= \psi_{(0)} + \lambda \psi_{(1)} + \lambda^2 \psi_{(2)} + \cdots \end{aligned} \tag{2.10}$$

By inserting eq. 2.9 and eq. 2.10 into the Schrödinger equation and after ordering the terms with the same λ parameters, one gets:

$$\begin{aligned} \hat{H}_0 |\psi_{(0)}\rangle &= E_{(0)} |\psi_{(0)}\rangle \\ \left(\hat{H}_0 - E_{(0)}\right) |\psi_{(1)}\rangle &= \left(E_{(1)} - \hat{V}\right) |\psi_{(0)}\rangle \\ \left(\hat{H}_0 - E_{(0)}\right) |\psi_{(2)}\rangle &= \left(E_{(1)} - \hat{V}\right) |\psi_{(1)}\rangle + E_{(2)} |\psi_{(0)}\rangle \end{aligned} \tag{2.11}$$

In the case of the second order perturbation theory, *i.e.*, MP2, the truncation of the perturbation expansion appears after the second order energy term $E_{(2)}$. By assuming that the perturbed wavefunctions are orthogonal to the zeroth order functions, *i.e.* $\langle \psi_{(0)} | \psi_{(i)} \rangle = \delta_{i0}$, the normalization $\langle \psi | \psi_{(0)} \rangle$ can be used to obtain the following expression for the energies up to the second order term:

$$
\begin{aligned}
E_{(0)} &= \left\langle \psi_{(0)} \middle| \hat{H}_0 \middle| \psi_{(0)} \right\rangle \\
E_{(1)} &= \left\langle \psi_{(0)} \middle| \hat{V} \middle| \psi_{(0)} \right\rangle \\
E_{(2)} &= \left\langle \psi_{(0)} \middle| \hat{V} \middle| \psi_{(1)} \right\rangle
\end{aligned}
\tag{2.12}
$$

Since the HF wavefunction $\left|\psi_{(0)}\right\rangle$ is an eigenfunction of $\hat{H}_0$, the zeroth-order energy reads

$$
E_{(0)} = \sum_a \epsilon_a \tag{2.13}
$$

With the definition of $\hat{V}$ (see eq. 2.8) in mind, the first-order energy is

$$
\begin{aligned}
E_{(1)} &= \left\langle \psi_{(0)} \middle| \hat{V} \middle| \psi_{(0)} \right\rangle \\
&= \left\langle \psi_{(0)} \middle| \sum_{\substack{i \\ j>i}} \frac{1}{|r_i - r_j|} \middle| \psi_{(0)} \right\rangle - \left\langle \psi_{(0)} \middle| \sum_i v^{HF}(i) \middle| \psi_{(0)} \right\rangle \\
&= \frac{1}{2} \sum_{\substack{i \\ j \neq i}} \langle ij| \, |ij\rangle - \sum_i \left\langle i \middle| v^{HF}(i) \right| i \rangle \\
&= -\frac{1}{2} \sum_{\substack{i \\ j \neq i}} \langle ij| \, |ij\rangle
\end{aligned}
\tag{2.14}
$$

The HF energy is then the sum of the zeroth- and first-order energies, and the first correction to the HF energy occurs, hence, in the second-order of the perturbation theory. The states $\left|\psi_{(1)}\right\rangle$ cannot be single excitations because of the Brillouin's theorem. Due to the two-particle nature of the perturbations, triply excited states do not mix with $\left|\psi_{(0)}\right\rangle$. Therefore, we are left with double excitations of the the form $|\psi_{ab}^{rs}\rangle$. Since the sum runs over all occupied space a and b and over all virtual space r and s, the second-order energy reads:

$$
E_{(2)} = \sum_{\substack{a \\ b>a}} \sum_{\substack{r \\ s>r}} \frac{\left| \left\langle \psi_0 \middle| \sum_i \sum_{j>i} |r_i - r_j|^{-1} \middle| \psi_{ab}^{rs} \right\rangle \right|^2}{\epsilon_a + \epsilon_b - \epsilon_r - \epsilon_s} = \sum_{\substack{a \\ b>a}} \sum_{\substack{r \\ s>r}} \frac{|\langle ab \,||\, rs\rangle|^2}{\epsilon_a + \epsilon_b - \epsilon_r - \epsilon_s} \tag{2.15}
$$

MP2 appears to be the simplest way to account for correlation energy. During the last few years, MP2 has been at the origin of new WFT development. Even though MP2 performs excellently for some types of interactions, such as hydrogen bonds, it provides an inadequate description of the weak intermolecular interactions.[33] More precisely, MP2 underestimates bond distances and overestimates interaction energies. Interaction of benzene dimer,[34] DNA base pairs and amino acids pairs,[35] prediction of metal-ligand bond dissociation energies and bond lengths[36] are examples of the poor MP2 behavior. S. Grimme suggested a cost-free modification known as SCS-MP2.[10] The latter considerably improved the description of correlation energy upon MP2. A few years later, a simple modification of the SCS-MP2, in which the same-spin component of the total MP2 energy was eliminated, was suggested by M. Head-Gordon *et al.*. The resulting SOS-MP2[37;38] lead to similar results to those obtained with SCS-MP2, with a considerable speed-up upon SCS-MP2. More recently, A. Tkatchenko *et al.* suggested a modification of the long-range part of the second-order energy by the use of a better C_6 coefficient.[39;40] Similary, in the MP2C[41;42] method, the dispersion part of the MP2 energy was replaced by that of the time-dependent DFT[43–45] (TD-DFT). In 2012, M. Head-Gordon *et al.* came

up with a attenuated Coulomb operator, leading to a range-separated MP2. Both MP2($erfc$) and MP2($terfc$), based on the $erfc$ and $terfc$ error functions, respectively, showed reduced deviations for the non-bonded interactions.[15] In parallel, tremendous efforts have been made to develop cost-effective MP2 code. In 1997 the resolution-of-identity (RI) approximation (see section III) has been introduced to solve the critical four-index two-electron repulsion integrals (ERI). The resulting RI-MP2[46–50] is suitable for large molecular systems, and provides a considerable speed-up. Other cost-effective methods have introduced the local *ansatz* and the truncation of the long-range correlation terms (LMP2,[51] Laplace transform MP2[52;53]).

C. Coupled cluster methods

The basic element of the coupled cluster[5–7] (CC) theory is a cluster expansion where one-body, two-body, three-body, *etc.*, clusters are the fundamental entities. The wavefunction *ansatz* of any CC method takes the following form:

$$\Psi = \mathrm{e}^{\hat{\mathrm{T}}}\Psi_0 \tag{2.16}$$

where Ψ_0 is a reference function, and $\hat{\mathrm{T}}$ is the excitation operator, which may be divided into various cluster terms $\hat{\mathrm{T}} = \hat{\mathrm{T}}_1 + \hat{\mathrm{T}}_2 + \hat{\mathrm{T}}_3 + ...$ The truncation of $\hat{\mathrm{T}}$ defines the CC method. Size-consistency is ensured by the disconnected cluster terms arising from the expansion of $\mathrm{e}^{\hat{\mathrm{T}}}$ in a power series. The first correction on the energy comes from the double-excitation. Therefore, CCD[54;55], where $\hat{\mathrm{T}} = \hat{\mathrm{T}}_2$, was the first CC method introduced. A few years later, R. J. Bartlett *et al.* developed CCSD[56], where $\hat{\mathrm{T}} = \hat{\mathrm{T}}_1 + \hat{\mathrm{T}}_1$, which provides results of semi-quantitative accuracy for a variety of molecules, it is not complete for many-electron systems.[29] Probably the most obvious strategy is to include higher excitation terms in the exponential *ansatz*. Towards a full CCSDT[57] model formulated and computationally implemented for the first time in 1987, R. J. Bartlett *et al.* developed the CCSDT-n[58;59] scheme, where higher values of n indicate fewer approximations to the CCSDT equations. In 1989, M. Head-Gordon *et al.* introduced the CCSD(T),[60] where the triple correction resulting from both single and double excitations, are non-iteratively treated via the triple correction formula (eq. 2.17).

$$\Delta\mathrm{E_T} = \sum_{\mathrm{t}}^{\mathrm{T}} \frac{|\langle\Psi\,|\mathrm{V}|\,\Psi_\mathrm{t}\rangle|^2}{(\mathrm{E}_0 - \mathrm{E}_\mathrm{t})} \tag{2.17}$$

where Ψ is the CCSD wavefunction, Ψ_t is the wavefunction composed by triples, $(\mathrm{E}_0 - \mathrm{E}_\mathrm{t})$ is the triple excitation energy using the Fock Hamiltonian (F), and $\mathrm{V} = \mathrm{H} - \mathrm{F}$. The performance of CCSD(T) is very close to the one of the full CCSDT method, but at considerably lower computational cost. Today, CCSD(T) is considered as the *gold standard* of computational chemistry

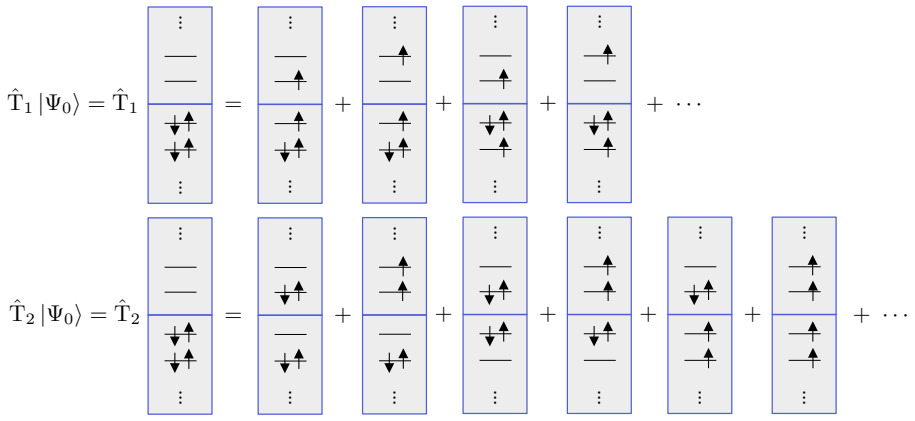


Figure 2.2: Schematic representation of $\hat{T}_2$ and $\hat{T}_1$ applied on the reference wavefunction Ψ_0.

D. Density functional theory

Density Functional Theory[8;9] (DFT) has become the most widely used quantum chemical method to account for electron correlation. The main reason for such large popularity is that correlated methods based on the $4N$-dimensional many-electron wavefunction scale poorly with system size (*e.g.* CCSD(T) scales as $O\left(N^7\right)$). In addition, they require high angular momentum basis functions to describe the electron-electron cusp,[61] whereas DFT, a formally exact theory based on one-electron density, depends on only three spatial coordinates and one spin coordinate. It is therefore computationally much simpler and, since the wavefunction is not explicitly modeled, the high angular momentum functions are much less important, allowing the usage of modest basis sets. It follows that DFT offers a favorable ratio between computational cost and accuracy.

The basis for DFT is the proof by Hohenberg and Kohn[8] that the ground state electronic energy is determined completely by the electron density ρ: there is an exact one-to-one mapping between ρ of a system and its energy. Further work from E. B. Wilson showed that (i) the integral of ρ defines the number of electrons, (ii) the cusps in ρ define the position of the nuclei, and (iii) the heights of the latter cusps define the corresponding nuclear charge.[62] The *only* problem being that, although it has been proven that each different density yields a different ground state energy, the functional connecting these two quantities still remains unknown. Therefore, the success of DFT critically depends on the quality of the exchange-correlation functional $E_{xc}[\rho]$, in which the interaction energy between electrons has been approximated. More precisely, $E_{xc}[\rho]$ is the result of the subtraction of the non-interacting kinetic energy, and the $E_{ne}[\rho]$ and $J[\rho]$ potential energy terms:

$$E_{xc}[\rho] = (T[\rho] - T_S[\rho]) + (E_{ne}[\rho] - J[\rho]) \tag{2.18}$$

The Jaccob's Ladder

The fundamental difficulty in DFT is that we do not know how to write the correct exchange-correlation energy, which includes exchange, correlation, and a small kinetic component. There are many categories of approximations, including:

- ▷ local density approximation (LDA)
- ▷ generalized gradient approximation (GGA)
- ▷ meta-GGAs
- ▷ hybrid functionals
- ▷ and meta-hybrid-GGAs

This hierarchy of methods has famously been termed by J. Perdew as the *Jaccob's Ladder*.[63] We now consider all these categories separately one at a time.

Local (Spin) Densiy Approximation. The local density approximation (LDA) and its extension to fermionic systems local spin density approximation (LSDA), are the first and easiest examples of approximations used in Kohn-Sham DFT. The general idea at their basis is (rather) simple: first the exchange-correlation energy per particle e_{xc} is computed for a homogeneous electron gas and then the global exchange-correlation energy is obtained for a generic system by weighting this quantity with probability $\rho(r)$, and integrating over all space.

$$E_{xc}^{LDA} = \int dr\ e_{xc}\left(\left(\rho(r)\right)\right)\rho(r) \tag{2.19}$$

The exchange-correlation energy per particle, e_{xc}, can be further divided into exchange and correlation contributions:

$$e_{xc}\left(\rho(r)\right) = e_x\left(\rho(r)\right) + e_c\left(\rho(r)\right) \tag{2.20}$$

The exchange part of a uniform electron gas is a functional of the density, and is easily obtained from geometrical considerations.

$$e_x^{LDA} = -\frac{3}{4}\left(\frac{3}{\pi}\rho(r)\right)^{1/3} \tag{2.21}$$

This expression leads to the well-known $\rho^{4/3}$ dependency of LDA. Analogous analytic expressions for the correlation part are not known, except for the two extreme cases of high- and low-density. In all the available correlation functionals, a parametrization of the accurate homogenous electron gas energies is obtained via quantum Monte Carlo simulations. The first simulation of homogenous electron gas was performed by D. M. Ceperley and B. J. Alder,[64] and various mathematical fits to their results are available in every DFT softwares.

Generalized Gradient Approximation. The LDA assumes a constant electron density, which is far from true in molecular systems. The natural next step is therefore to introduce additional information about the density gradient. The resulting functionals are termed Generalized Gradient Approximation (GGA). By introducing the dimensionless reduced gradient variable s the correct exchange scaling is always achieved.[65]

$$s(r) = \frac{\nabla s(r)}{s^{4/3}(r)} \tag{2.22}$$

Integration of the per-particle exchange functional leads to the final form of the GGA exchange functional which reads:

$$E_x^{GGA}[\rho] = \int dr\ \rho^{4/3}(r) F\left(s(r)\right) \tag{2.23}$$

where $F\left(s(r)\right)$ can be seen as a gradient expansion. There are many GGAs in literature, and approximations can be classified as either semi-empirical (*i.e.* fitted to experimental data, such as BLYP,[66;67] OLYP,[68;67] HCTH[69]) or purely theoretical (*i.e.* determined by satisfying exact conditions, such as PW91,[70] PBE[71]).

Meta-GGAs. Following the LDA and the GGA, the logical next step is to introduce higher derivatives of the density into the functional, namely the density Laplacian $\nabla^2\rho$ and the kinetic energy density $\tau = \sum_i |\nabla\phi_i|^2$, the integral of which over space is the non-interacting kinetic energy. The increased number of parameters in these functionals allows more exact conditions to be satisfied and, as a result, meta-GGAs are able to provide an improved description on many different systems. However, for molecules, they do not offer any significant improvement over other categories of functionals and are consequently not widely used in chemistry.

Hybrids. The starting point of the hybrid functionals is the adiabatic connection,[72;73] which links the non-interacting electron-electron repulsion to the fully interacting case by a smooth function of a parameter λ that runs from zero (no interaction) to one (full interaction):

$$E_{xc} = \int_0^1 d\lambda\ \left\langle \Psi_\lambda \left| \hat{V}_{ee} \right| \Psi_\lambda \right\rangle - J\left[\rho\right] = \int_0^1 d\lambda\ W_\lambda \tag{2.24}$$

In 1993, A. D. Becke attempted an integration of linear form of E_λ (in eq. 2.24) which lead to the *half-and-half* functional.[74] The latter suggested a semi-empirical mixing of the *non-local* HF exchange to the *local* exchange-correlation functional:

$$E_{xc} = C_x E_x^{GGA} + (1 - C_x) E_x^{HF} + E_c^{GGA} \tag{2.25}$$

As with GGAs, there are numerous hybrid functionals in literature, which can be categorised as either semi-empirical (*e.g.* B3LYP,[74] B97-3[75]) or purely theoretical (*e.g.* PBE0[76]).

Meta-hybrid-GGAs. Hybrid functionals are successful in the description of short-range electron-electron interactions. As a result, hybrid functionals perform very well in the evaluation of atomisation energies, ionization potentials, electron affinity and bond-lengths. However, they perform poorly for the evaluation of Rydberg states, and charge transfers. The reason for this failure is the inadequate description of the long-range exchange interaction as an increased amount of exact exchange is required. The Coulomb-attenuation or range-separation approach proposed by T. Yanai[77] solves this problem. The latter approach splits the electron-electron repulsion operator into a long- and a short-range exchange:

$$\frac{1}{|r_1 - r_2|} \equiv \underbrace{\frac{[\alpha + \beta \mathrm{erf}(\mu\,|r_1 - r_2|)]}{|r_1 - r_2|}}_{\mathrm{LR}} + \underbrace{\frac{1 - [\alpha - \beta \mathrm{erf}(\mu\,|r_1 - r_2|)]}{|r_1 - r_2|}}_{\mathrm{SR}} \tag{2.26}$$

The first component calculates the long-range exchange (LR), while the second component gives the short-range exchange (SR). The amount of exchange increases as r_{12} increases in as much as the functionals behave like a standard hybrid with an amount α of exact exchange at short range and get closer to HF (plus correlation) at long-range. The parameter μ controls the rate of attenuation.

Stairway to heaven.

Following the classifications from Perdrew, J. Klimeš and A. Michaelides suggested in 2012 the *stairway to heaven*,[31] a classification scheme of DFT-based dispersion scheme on the level of approximation each method makes in obtaining the long range dispersion interactions.

Standard DFT methods. The first rung of the *stairway to heaven* is occupied by approaches that give incorrect shapes of binding curves and underestimates the binding of well-separated molecules. The total energy of the system, E_{tot}, is simply the DFT energy, E_{DFT}.

$$E_{tot} = E_{DFT} \tag{2.27}$$

Simple C_6 empirical correction. The basic requirement for any DFT-based dispersion scheme should be that it yields reasonable $^1/_{r^6}$ asymptotic behavior for the interaction of particles in the gas phase, where r is the distance between the particles. A simple approach for achieving this is to add an empirical term accounting for the missing long range attraction. The total energy then reads

$$E_{tot} = E_{DFT} + E_{disp.} \tag{2.28}$$

where $E_{disp.}$ is the dispersion interaction computed as:

$$E_{disp.} = -\sum_{A,B} \frac{C_6^{AB}}{r_{AB}^6} f_{AB}\,(r_{AB}, A, B) \tag{2.29}$$

where the dispersion coefficients C_6^{AB} depend on the elemental pairs A and B. Since the dispersion correction diverges at short inter-atomic separations and the dispersion correction part is damped by

the function $f_{AB}(r_{AB}, A, B)$ which is equal to one for large r_{AB} and decreases E_{disp} to zero or to a constant for small values of r_{AB}. These methods are generally termed "DFT-D".

Environment-dependant C_6 correction. An inconvenient of the "DFT-D" schemes is that the dispersion coefficients are predetermined and constant quantities. Therefore the same coefficient will be assigned to an element regardless of its oxidation or hybridization state. The errors introduced by this approximation can be large, *e.g.* the carbon C_6 coefficients can differ by almost 35% between the sp and sp3 hybridized states.[78] Hence, the emergence of methods where the C_6 coefficients vary with the environment of the atom has been a very welcome development.

Non-local DFT. The approaches associated with the fourth rung of the *stairway to heaven* do not rely on external input parameters but obtain the dispersion interaction directly from the electron density. The methods have been termed non-local correlation functionals since they add non-local (i.e., long range) correlations to local or semi-local correlation functionals. The non-local correlation energy E_c^{nl} is defined as a double space integral over the electron density and includes a classical Coulomb-like interaction kernel (see eq. 2.30).

$$E_c^{nl} = \int\int dr_1 dr_2 \; \rho(r_1) O\left(\frac{1}{|r_1 - r_2|^6}\right) \rho(r_2) \tag{2.30}$$

Within this approach the exchange-correlation energy reads:

$$E_{xc} = E_x^{GGA} + E_c^{LDA} + E_c^{nl} \tag{2.31}$$

The double-hybrid scheme. In its more typical formulation, the double-hybrid (DH) scheme mixes an exact exchange term with the DFT exchange functional as simple hybrids, but adds a perturbational correlation term to the DFT correlation in the basis of the Kohn-Sham orbitals.[10;11]

$$E_{xc} = C_x E_x^{HF} + (1 - C_x) E_x^{GGA} + C_c E_c^{GGA} + (1 - C_c) E_c^{MP2} \tag{2.32}$$

The first DH of this form was the B2PLYP of S. Grimme,[79] with other examples being the general purpose B2GP-PLYP2[80] and the long-range corrected ωB97X-2.[81] Flexibility was provided by setting a different weight to the same-spin and opposite-spin MP2 correlation contributions, in the spirit of SCS-MP2.[10] This lead to the dispersion-corrected spin-component-scaled DH[12] (DSD) DFTs with an exchange-correlation functional:

$$E_{xc} = C_x E_x^{HF} + (1 - C_x) E_x^{GGA} + C_c E_c^{GGA} + C_{c,o} E_{c,o}^{MP2} + C_{c,p} E_{c,p}^{MP2} + C_6 E_6^{disp} \tag{2.33}$$

The DSD functionals were implemented in the GAMESS software[14] in this work, as a first approach to density functional implementations (*c.f.* section VII hereafter).

E. Note on the performance of the available quantum chemical methods

Each of the above-mentioned quantum chemical methods has a certain level of accuracy closely related to its level of theory (*i.e.* the level of approximations), and ultimately to the computational cost.

If one can suggest a general *step ladder* aiming at a general hierarchy of methodologies, the latter would start with Hartree-Fock which is surprisingly accurate considering the simplicity of its *ansatz*. HF works specially well for the main group elements and for formally d^0 transition metals.[28] Obviously HF fails to describe electron delocalization and correlation effects. As illustrated in Fig. 2.4, HF scales as $O\left(N^4\right)$ with the system size. In addition, small basis sets, such as, *e.g.*, split-valence or

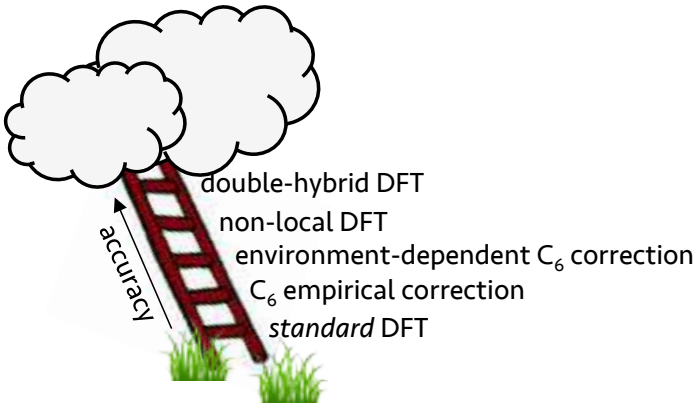


Figure 2.3: Schematic representation of the stairway to heaven.

double-ζ quality, are usually sufficient for routine treatment, as detailed in section V of this Chapter. DFT would come next with a scaling close to $O\left(N^3\right)$. Amongst all methodologies, DFT provides the best accuracy over computational cost ratio. Although DFT often provides greater accuracy in bond energies and reaction barriers for less computational effort,[82] a range of progressively more sophisticated electron correlated methods, such as MP2 ($O\left(N^5\right)$), is often superior for intermolecular interactions.[83] A pertubative treatment of the zeroth order HF wavefunction typically cuts down the HF error by *ca.* 60 %. However, MP2 requires larger basis set than HF (see section V). Aiming at high accuracy and reliability coupled cluster methods with CCSD (scaling as $O\left(N^6\right)$) and CCSD(T) (scaling as $O\left(N^7\right)$)[84] are of benchmark quality, but are generally not applicable to large systems (*i.e.* above 800 basis functions), although this challenge is being addressed by ongoing developments in explicitly correlated and local correlation methods.[85;86]

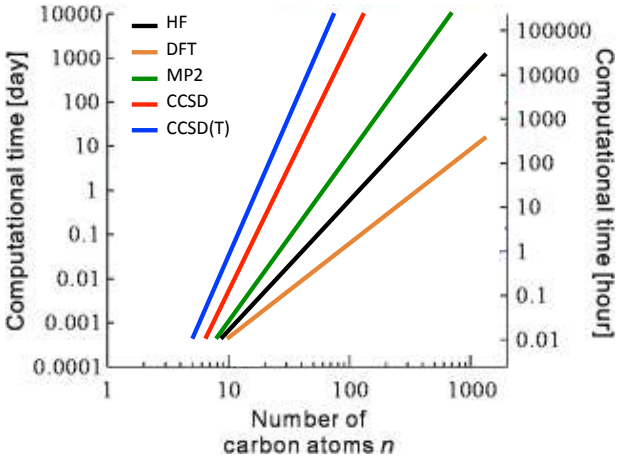


Figure 2.4: Approximate computational time scaling of various computational methods with the number of carbon atoms N: Hartree-Fock (HF) scales with $O\left(N^4\right)$, density functional theory (DFT) with $O\left(N^3\right)$, Møller-Plesset second order perturbation theory (MP2) with $O\left(N^5\right)$, and coupled-cluster (CCSD(T)) with $O\left(N^7\right)$.

Hereafter, coupled cluster is used as state-of-the-art method which is typically the level chosen against which the performance of new and/or existing methodologies are compared, as the gold standard.

II Integrals over Cartesian Gaussian in *ab initio* calculations

The great success of the Gaussian type orbitals (GTOs), with the following general formulation

$$\chi_i^{PGF}(r) \equiv (x - I_x)^{i_x} (y - I_y)^{i_y} (z - I_z)^{i_z} \exp\left[-\alpha |r - A|^2\right] \tag{2.34}$$

is due to the fact that all fundamental integrals (see eq. 2.35), *i.e.* overlap, kinetic-energy, electron-repulsion, nuclear-attraction, and anti-Coulomb integrals, are easily evaluated analytically at a tolerable computational cost. The main reason for this efficiency is the Gaussian product theorem (GPT). In eq. 2.34, the primitive Gaussian function χ_i^{PGF}, centered in $A = (A_x, A_y, A_z)$, has an angular momentum $a = (a_x, a_y, a_z)$, and an exponent α. Most of the matrix elements arising from computing the SCF energy and its derivatives with respect to nuclear motion can be written in terms of integrals of the general form, with $F(x)$ being a very simple function (*e.g.* $x = {}^1/_x$)

$$I = \int\int d\mathbf{r_1} d\mathbf{r_2} \chi_a(\mathbf{r_1}) \chi_b(\mathbf{r_1}) F(|(\mathbf{r_1}) - (\mathbf{r_2})|) \chi_c(\mathbf{r_2}) \chi_d(\mathbf{r_2}) \tag{2.35}$$

A. Integrals in GAMESS

Fig. 2.5 shows the flow diagram of the subroutines involved in the computation of the integrals for a single-point MP2 calculation. The interest for MP2 is not innocent: the implementation of an attenuated MP2 scheme, as reported in section X of the present Chapter, involves modifications of the Coulomb operator, and, thus, the integrals and their respective algorithms.

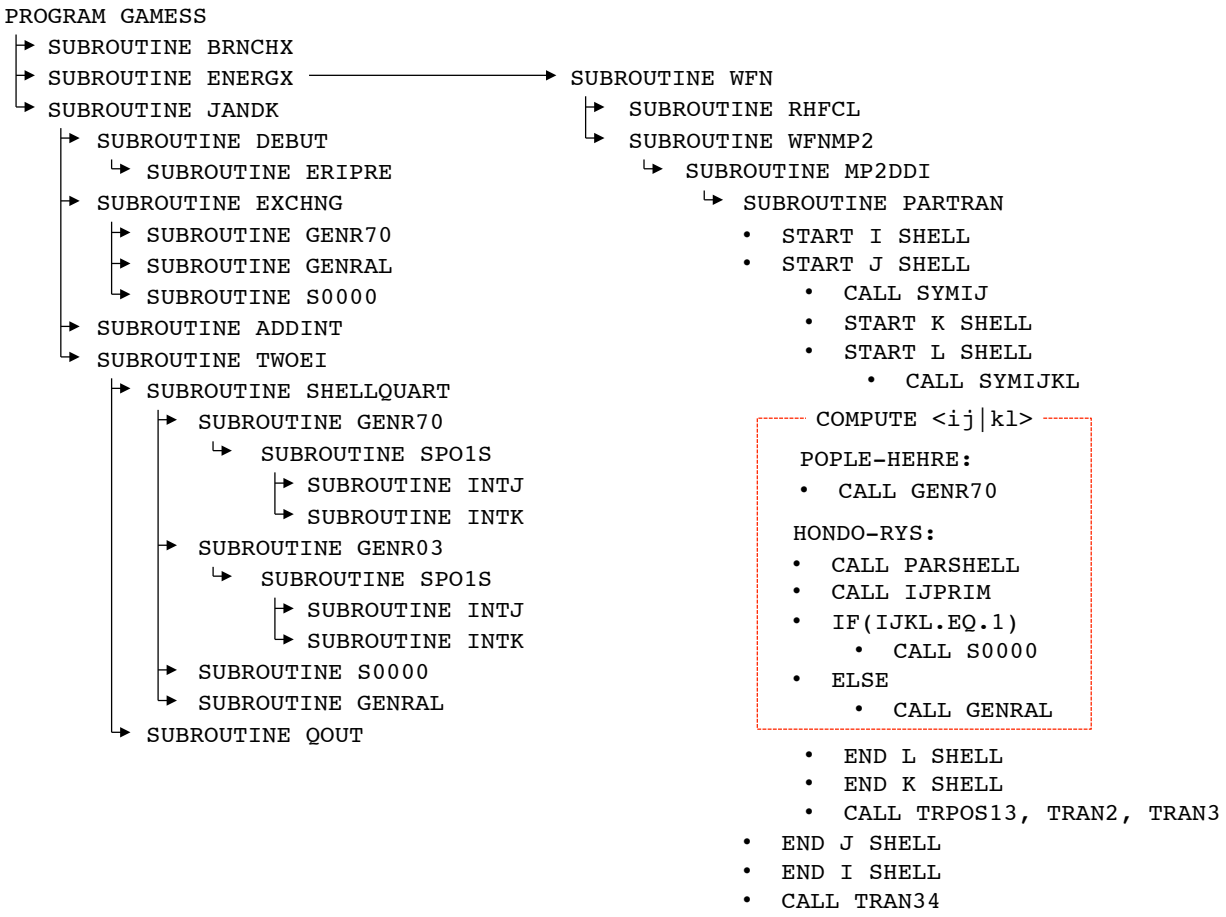


Figure 2.5: Flow diagram showing the subroutines involved in the computation of the integrals for a single-point MP2 calculation.

The Pople-Hehre algorithm[87] is exceptionally efficient for computing integrals I with highly contracted s and p functions. It uses information common to a shell of basis functions and computes the desired integrals by accumulating combinations over several sets of Cartesian axes. Moreover, the use of different system of coordinate axes fully participates in the computational efficiency as many primitive integrals vanish by symmetry argument. Although this algorithm is impressively performant on s and p functions, it founders when applied to functions with higher angular momentum: the cost generated by the rotation steps from one coordinate axis to another being the reason. Consequently, a different algorithm is used for such class, referred to as the Hondo-Rys algorithm. The latter technique is a numerical integration using the Rys polynomials.[88–90]

The method of Rys quadrature offers remarkable advantages for the evaluation of molecular integrals over Gaussian basis functions with high angular momentum. The idea is to re-express the integrals by obtaining a polynomial approximation valid over a specific range of the parameter. Computationally stable procedure for the evaluation of the integral is to generate the so-called fundamental integral and derive all other integrals by a downward recursion.

In the current version of GAMESS, users can easily select amongst different algorithm flavors for computing the integrals. The branching is ensured by the `INTTYP` keyword. The following algorithm produces equally accurate results, and are, therefore, mostly used for debugging purposes:

- ▷ `BEST` uses the fastest integral code available for any particular shell quartet. This is the default value for `INTTYP`. The choice order based on the shell quartet is first `ROTAXIS`, then `ERIC` and finally `RYSQYAD`.
- ▷ `ROTAXIS` uses only rotated axis codes for s,p,L shells, and the Rys quadrature for other type of shells.
- ▷ `RYSQUAD` uses the Rys quadrature for everything.

In addition, the `ERIC` precursor algorithm (not mentioned in this work) can also be used by setting `INTTYP=ERIC`.* In section X, particular emphasis on modifying and adapting the `ROTTAXIS` and `RYSQUAD` algorithms to new methodologies is provided.

*more information is found in the GAMESS input manual.

III The resolution-of-identity approach

In this section, the basic resolution of identity (RI) formalism is presented, synonymously referred to as density fitting technique.[91] The ultimate goal behind the different versions of RI[16;92–97] is to reduce the cost of the expensive (in CPU time, memory, and disk storage) four-orbital two-electron Coulomb integrals

$$\langle ij|kl\rangle = \int\int dr_1 dr_2 \; \frac{\phi_i(r_1)\phi_j(r_1)\phi_k(r_2)\phi_l(r_2)}{|r_1 - r_2|} \tag{2.36}$$

which is a major bottleneck for the approaches beyond DFTs (see section I). The idea behind the RI-approximation is to represent pair products of atomic basis functions $\chi_i(r)\chi_j(r)$ (where $\phi_i = \sum_A C_{iA}\chi_{iA}$ eq. 2.36) in terms of auxiliary basis functions, such as

$$\rho_{ij}(r) \equiv \chi_i(r)\chi_j(r) \;\approx\; \tilde{\rho}_{ij}(r) \equiv \sum_{\mu=1}^{N_{aux}} C_{ij}^{\mu} P_\mu(r) \tag{2.37}$$

where μ labels the auxiliary basis functions (P_μ), C_{ij}^{μ} are the expansion coefficients, and $\rho_{ij}(r)$, and $\tilde{\rho}_{ij}(r)$ denote pair product of basis function and their approximate expansion in auxiliary basis functions, respectively. Inserting eq. 2.37 into eq. 2.36 leads to the following expression:

$$\langle ij|kl\rangle \approx \sum_{\mu}\sum_{\nu} C_{ij}^{\mu} \int\int dr_1 dr_2 \; \frac{p_\mu(r_1)p_\nu(r_2)}{|r_1 - r_2|} C_{kl}^{\nu} = \sum_{\mu,\nu} C_{ij}^{\mu} \langle\mu|\nu\rangle C_{kl}^{\nu} \tag{2.38}$$

A. RI-approximation in GAMESS

Among the different versions of RI to determine the expansion coefficients C_{ij}^{μ},[16;92–97] GAMESS uses the so-called *RI-V* method from Whitten,[16] which minimizes the RI error of the four-center integrals themselves:

$$\delta I_{ij,kl} = \langle\tilde{\rho}_{ij}|\tilde{\rho}_{kl}\rangle - \langle\rho_{ij}|\rho_{kl}\rangle \tag{2.39}$$

The minimization of $\delta I_{ij,kl}$ is achieved by independently minimizing the self-repulsion of the basis pair density residuals. It leads to the following decomposition of the four-center electron repulsion integrals:

$$\langle ij|kl\rangle \approx \sum_{\mu}\sum_{\nu} \langle ij|\mu\rangle \langle\mu|\nu\rangle^{-1} \langle\nu|kl\rangle \tag{2.40}$$

Hence, the expensive four-center integrals reduces to the much cheaper three- and two-center integrals. The performance of such approximation is intensively studied in section VB. It is also at the heart of the cost-effective methods developed hereafter and reported in section IX of this Chapter.

GAMESS user can select the RI-approximation to solve the four-index two-electron integrals of MP2: `CODE=RIMP2`.

IV Validation data sets

Validation data sets, also referred to as test sets, are of utmost importance in assessing the performance of existing methods and in the process of optimizing and benchmarking new methods.

Despite the large number of existing validation sets, we were in the need to develop new sets categorized in specific interaction energy range and type. As a matter of fact, even though non-covalent interactions result from the dynamic electron correlation, they can further be divided into various subclasses. Typically hydrogen-bonds, a specific type of interactions invloving dipole-dipole interactions between a partially positive hydrogen atom and a highly electronegative, partially negative oxygen, nitrogen, sulfur, or fluorine atom, is central to biochemistry as they are highly specific and directional. Halogen bonds are similar to hydrogen-bonds with the difference that a halogen atom is involved in the interacting players. This type of interaction are also referred to as charge transfer interactions. Although charge-transfer does not play a decisive role in chemical systems, they should be properly considered. All non-covalent interaction involving a dipole is known under the so-called van der Waals (vdW) forces. Dipole-dipole interactions are electrostatic interactions between permanent dipoles in molecules. These interactions tend to align the molecules to increase attraction. A dipole-induced dipole interaction, best known under the Debye forces, is due to the approach of a molecule with a permanent dipole to another non-polar molecule with no permanent dipole. This proximity causes the electrons of the non-polar molecule to be polarized toward or away from the dipole, inducing a dipole. London dispersion forces are the weakest type of non-covalent interactions. They are also known as induced dipole-induced dipole interactions and are present between all molecules, even those which inherently do not have permanent dipoles. They are caused by the temporary repulsion of electrons away from the electrons of a neighboring molecule, leading to a partially positive dipole on one molecule and a partially negative dipole on another. Last but not least π-effects are associated with the interaction between the π-orbitals of a molecular system.[98;99]

In this regard, seven new validation data sets were created to cover a broad range of non-covalent interactions and in particular, to include the classes described earlier: hydrogen-bond, charge-transfer, dipole-dipole, vdW forces though π-effects, interaction between rare gases atoms, and alkane dimerization. Overall, 66 chemical systems with correlation energy ranging from 18.6 kcal/mol to 0.02 kcal/mol were selected. They are summarized in Fig. 2.6, and their respective values are reported in the tables of their corresponding subsections.

By assessing *ab initio* methods on specific validation sets such as the one designed in this work, provides a solid understanding of their performance and gives a good feeling of which methods should be used to describe chemical systems with specific stabilizing contributions. A powerful usage of such approach is illustrated in the sections VII and VIII of this Chapter.

HB9 is a data set which includes nine complexes governed by hydrogen bonds. The complexes are depicted in Fig. 2.6 (A) and their corresponding interaction energies are reported in table 2.1.

Systems	E_{int}
$H_2O \cdots H_2O$	-5.006[100]
$H_2O \cdots NH_3$	-6.493[100]
$NH_3 \cdots NH_3$	-3.137[100]
$HCONH_3 \cdots HCONH_3$	-15.96[99]
$HCOOH \cdots HCOOH$	-18.61[99]
$HF \cdots MeNH_2$	-14.32[101]
$HF \cdots MeOH$	-9.59[101]
$HCN \cdots HCN$	-4.745[100]
$HF \cdots HF$	-4.581[100]

Table 2.1: Theoretical interaction energies E_{int} in kcal/mol for the HB9 test set.

CT7/04 is a data set designed by D. G. Truhlar *et al.*[102;103] It includes seven complexes governed by charge transfer interactions. The complexes are depicted in Fig. 2.6 (B) and their corresponding interaction energies are reported in table 2.2.

Systems	E_{int}
$H_2O \cdots ClF$	-5.36
$NH_3 \cdots ClF$	-10.62
$NH_3 \cdots Cl_2$	-4.88
$NH_3 \cdots F_2$	-1.81
$C_2H_2 \cdots ClF$	-3.81
$C_2H_4 \cdots F_2$	-1.06
$HCN \cdots ClF$	-4.86

Table 2.2: Theoretical interaction energies E_{int} in kcal/mol for the CT7 test set.

DI9 is a data set which includes nine complexes governed by dipole interactions. The complexes are depicted in Fig. 2.6 (F) and their corresponding interaction energies are reported in the table 2.3.

Systems	E_{int}
$CH_3Cl \cdots HCl$	-3.550[102;103]
$H_2S \cdots H_2S$	-1.660[102;103]
$CH_3F \cdots CH_3F$	-1.648[101]
$CH_3Cl \cdots CH_2O$	-1.170[101]
$CH_3Cl \cdots CH_3Cl$	-1.338[101]
$CH_3OH \cdots CH_3F$	-3.893[101]
$HCl \cdots H_2S$	-3.350[102;103]
$CH_3SH \cdots HCN$	-3.590[102;103]
$CH_3SH \cdots HCl$	-4.160[102;103]

Table 2.3: Theoretical interaction energies E_{int} in kcal/mol for the DI9 test set.

ADIM5 is a data set which is governed by dispersion interactions, via the dimerization process of five alkanes. The complexes are depicted in Fig. 2.6 (D) and their corresponding interactions energies are reported in table 2.4. The chemical systems were first optimized at the MP2/Def2-TZVPPD[104] level of theory in the D_{3d}, D_2, and C_{2h} symmetry point groups (see Fig. 2.6 (D) for more information). Then, single point calculations were performed with the Def2-QZVPD basis set on the converged geometry.

Systems	E_{int}
CH_4 dimer	-0.530[105]
C_2H_6 dimer	-1.353[105]
C_3H_8 dimer	-2.048[105]
C_4H_{10} dimer	-2.971[105]
C_5H_{12} dimer	-3.922[105]

Table 2.4: Theoretical interaction energies E_{int} in kcal/mol for the ADIM5 test set.

IDISP4 is a data set which covers internal dispersion interactions of two alkanes: butane and pentane. Each of them were optimized in three different conformations at the B97-D[106]/def2-QZVPD level of theory, within their corresponding symmetry point group, *i.e.*, C_1, C_2, C_{2v} and C_{2h}. The complexes are depicted in Fig. 2.6 (C) and their corresponding interactions energies are reported in table 2.5. Effects of thermal correction on the strength of internal dispersion were investigated and were found to be negligible (*ca.* 0.08 kcal/mol difference).

Systems	E_{int}
anti $\rightarrow$ gauche	+0.67[107]
anti $\rightarrow$ syn	+3.95[107]
anti-anti $\rightarrow$ anti-gauche	+0.618[108]
anti-anti $\rightarrow$ gauche-gauche	+0.940[108]

Table 2.5: Experimental interaction energies E_{int} in kcal/mol for the IDISP4 test set.

PPS11 is a data set which includes 11 complexes governed by $\pi - \pi$ interactions. The complexes are depicted in Fig. 2.6 (G) and their corresponding interaction energies are reported in table 2.6.

Systems	E_{int}
$C_2H_2 \cdots C_2H_2$	-1.537 [100]
$C_2H_2 \cdots C_2H_2$	+1.074 [100]
$C_2H_2 \cdots C_2H_2$	-1.34 [102;103]
$C_2H_4 \cdots C_2H_2$	+0.784 [100]
$C_2H_4 \cdots C_2H_2$	-1.53 [99]
$C_2H_4 \cdots C_2H_4$	-1.11 [100]
$C_2H_4 \cdots C_2H_4$	+0.898 [100]
$C_2H_4 \cdots C_2H_4$	-1.42 [102;103]
$C_6H_6 \cdots C_6H_6$	-2.78 [102;103]
$C_6H_6 \cdots C_6H_6$	-1.81 [102;103]
$C_6H_6 \cdots C_6H_6$	-2.74 [102;103]

Table 2.6: Theoretical interaction energies E_{int} in kcal/mol for the PPS11 test set.

RG21 is a data set which includes 21 complexes governed by weak interactions, via the dimerization of rare gases. The complexes are depicted in Fig. 2.6 (E) and their corresponding interaction energies are reported in table 2.7.

Systems	E_{int}	Systems	E_{int}
He $\cdots$ He	-0.022 [109;110]	Ar $\cdots$ Xe	-0.375 [109;110]
He $\cdots$ Ne	-0.041 [109;110]	Kr $\cdots$ Kr	-0.400 [109;110]
He $\cdots$ Ar	-0.057 [109;110]	Kr $\cdots$ Xe	-0.464 [109;110]
He $\cdots$ Kr	-0.057 [109;110]	Xe $\cdots$ Xe	-0.561 [109;110]
He $\cdots$ Xe	-0.054 [109;110]	He trimer	-0.061 [111]
Ne $\cdots$ Ne	-0.084 [109;110]	Ne trimer	-0.239 [111]
Ne $\cdots$ Ar	-0.134 [109;110]	Ar trimer	-0.850 [111]
Ne $\cdots$ Kr	-0.142 [109;110]	$CH_4 \cdots$ Ne	-0.220 [103]
Ne $\cdots$ Xe	-0.147 [109;110]	$CH_4 \cdots$ Ar	-0.405 [100]
Ar $\cdots$ Ar	-0.285 [109;110]	$C_2H_4 \cdots$ Ar	-0.364 [100]
Ar $\cdots$ Kr	-0.361 [109;110]		

Table 2.7: Experimental and theoretical interaction energies E_{int} in kcal/mol for the RG21 test set.

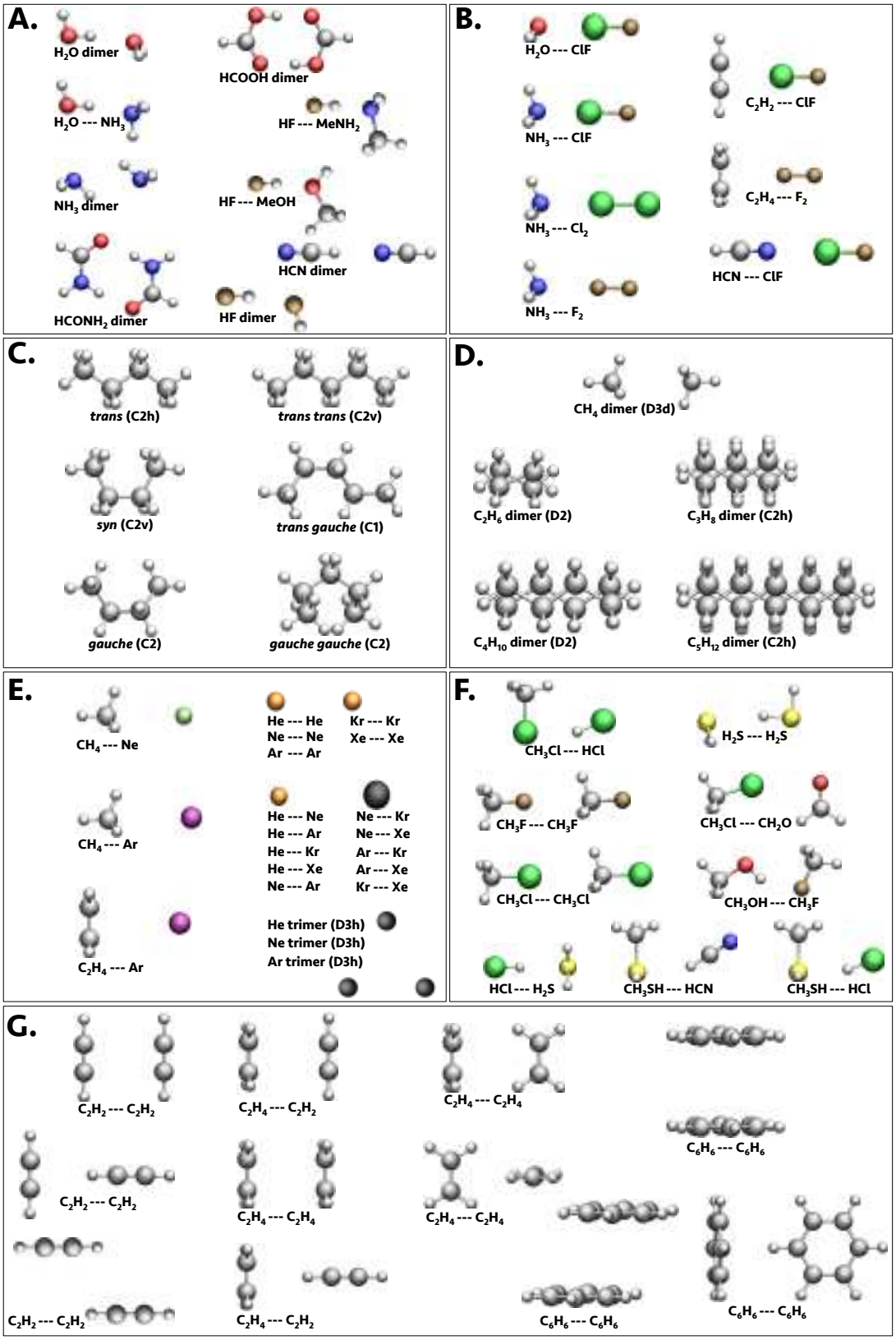


Figure 2.6: The seven different validation sets used to assess the performance of the computational methods. The sets cover hydrogen bonds (A), charge transfer interactions (B), internal dispersion forces (C), weak interactions – via alkane dimerization process (D), and via rare gas dimerzation/trimerization process (E) – dipole interactions (F), and $\pi - \pi$ interactions (G).

V Basis set considerations

In the present section an exhaustive basis set convergence study is carried out on the strongest and on the weakest interacting complexes of each data set, expected ADIM4 from which only CH_4 was included. The resulting set is referred to as a *reduced data set* and is displayed in figure 2.7. The choice of the basis set is of utmost importance when designing high accuracy quantum chemical calculations. In section A., a wide range of basis sets, from [3s2p1d] to [8s7p6d5f4g3h2i] contracted functions are benchmarked against the complete basis set (CBS) limit, computed by means of several CBS extrapolation procedures. In section B., the assessment of the auxiliary basis set used for the resolution-of-identity approximation is discussed.

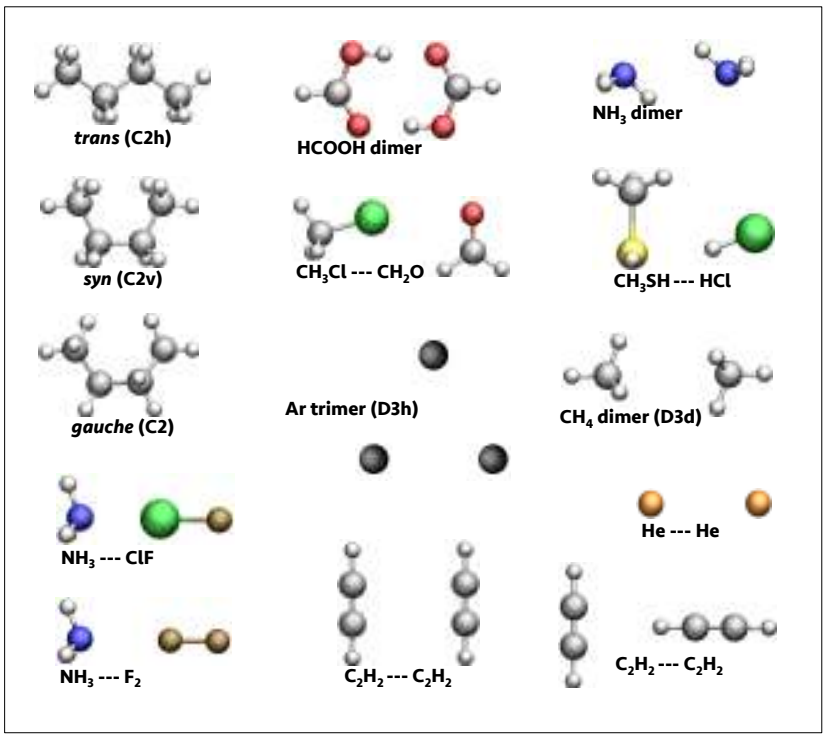


Figure 2.7: The reduced data set contains 13 complexes, including $He\cdots He$, $Ar\cdots Ar\cdots Ar$ (RG21), $NH_3\cdots ClF$, $NH_3\cdots F_2$ (CT7/04), $HCOOH\cdots HCOOH$, $NH_3\cdots NH_3$ (HB9), $CH_4\cdots CH_4$(ADIM4), $CH_3SH\cdots HCL$, $CH_3Cl\cdots CH_2O$ (DI9), *trans* (C2h), *syn* (C2v), *gauche* (C2) (IDISP4), $CH_2\cdots CH_2$, and $CH_2\cdots CH_2$ (PPS11).

A. Main basis set

Truncation of the one-electron space to a finite basis introduces a severe approximation in solving the Born-Oppenheimer electronic Schrödinger equation. Hence, any additional artifact arising from a poor basis set behavior has to be eliminated. Three families of standard basis sets are investigated in the present case: (i) the augmented correlation-consistent aug-cc-pVnZ (n=D,T,Q,5,6) developed by T. H. Dunning *et al.*,[112–117] (ii) the augmented correlation-consistent with tight diffuse function aug-cc-pCVnZ (n=D,T,Q) developed by T. H. Dunning *et al.*,[118;119] and (iii) the Def2-style basis set developed by R. Ahlrichs *et al.*.[120;121;104;122] A description of these basis sets, including their composition and their highest angular momentum functions L_{max} is found in table 2.8.

Herein, calculations were performed on the reduced set (Fig. 2.7) at the PBE,[71;123] DSD-PBEPBE,[13]*

*The implementation of the DSD-PBEPBE is detailed in section VII

Basis set	L_{max}	Contracted functions		Primitive *sp*-functions	
		1-Row	2-Row	1-Row	2-Row
aug-cc-pVDZ	2	4s3p2d	5s4p2d	18s4p	35s16p
aug-cc-pVTZ	3	5s4p3d2f	6s5p3d2f	19s6p	42s17p
aug-cc-pVQZ	4	6s5p4d3f2g	7s6p4d3f2g	22s7d	43s20p
aug-cc-pV5Z	5	7s6p5d4f3g2h	8s7p5d4f3g2h	25s9p	53s21p
aug-cc-pV6Z	6	8s7p6d5f4g3h2i	9s8p6d5f4g3h2i	28s11p	54s24p
aug-cc-pCVDZ	2	5s4p2d	6s5p3d	19s6p	36s17p
aug-cc-pCVTZ	3	7s6p4d2f	8s7p5d3f	21s8p	44s19p
aug-cc-pCVQZ	4	9s8p6d4f2g	10s9p7d5f3g	25s10p	46s23p
Def2-SVP	2	3s2p1d	4s3p1d	7s4p	10s7p
Def2-SVPD	2	4s3p2d	5s4p2d	8s5p	11s8p
Def2-TZVP	3	5s3p2d1f	5s5p2d1f	11s6p	14s9p
Def2-TZVPD	3	6s4p3d1f	6s6p3d1f	12s7p7	15s10p
Def2-TZVPP	3	5s3p2d1f	5s5p3d1f	11s6p	14s9p
Def2-TZVPPD	3	6s4p3d1f	6s6p4d1f	12s7p	15s10p
Def2-QZVP	4	7s4p3d2f1g	9s6p4d2f1g	15s9p	20s14p
Def2-QZVPD	4	8s5p4d2f1g	10s7p5d2f1g	16s10p	21s15p
Def2-QZVPP	4	7s4p3d2f1g	9s6p4d2f1g	15s9p	20s14p
Def2-QZVPPD	4	8s5p4d2f1g	10s7p5d2f1g	16s10p	21s15p

Table 2.8: Composition of the basis sets in terms of contracted and primitive basis functions. L_{max} is the highest angular momentum function of a particular set.

and MP2[4] level of theory, establishing convergence of the interaction energy with respect to increasing the basis set size. The choice of the computational methodologies shows the convergence of the density-based methods, the correlated models, and the mixed DSD-DFT. The N-electron methods combined with the T. H. Dunning n-tuple augmented correlation consistent aug-cc-pVnZ basis set (n=D,T,Q,5,6) allow extrapolation to the CBS limit and gives information on the intrinsic errors of the different N-electron models, enabling the investigation of the accuracy of standard basis sets.

The aug-cc-pVnZ set is particularly well suited for an extrapolation to the CBS limit, since with each extension of the basis set, new functions are added that make similar contribution to the energy. For example the [4s3p2d] aug-cc-pVDZ basis for the 1-Row (see table 2.8) contains a set of Hartree-Fock orbitals, augmented with one set of correlating functions for each occupied orbitals, and with one set of diffuse functions. Proceeding to the [5s4p3d2f] aug-cc-pVTZ basis, an additional set of correlating functions of each angular momentum along with a set of higher angular-momentum functions are included. This process is continued for larger sets, making sure that the correlating functions added at each step make similar contributions to the energy. Thus, given the hierarchical sequences of the Dunning aug-cc-pVnZ basis sets, a systematic improvement of the property f (the total energy in this particular case, see eq. 2.41) is obtained when climbing the steps towards the completeness of the set.

Based on numerical evidences, and following earlier work by D. Feller,[124;125] the relationship between the highest angular momentum L_{max} (see table 2.8) and f_{CBS} has been established to fit a three-point exponential formula of the form (eq. 2.41)

$$f\left(L_{max}\right) = f_{CBS} + A\exp\left(-\alpha L_{max}\right) \tag{2.41}$$

or (eq. 2.42)

$$f\left(L_{max}\right) = f_{CBS} + A\left(L_{max}+1\right)\exp\left(-\alpha\sqrt{L_{max}}\right) \tag{2.42}$$

Note that f_{CBS} is the value of the property f at the CBS limit.

Later, alternative fitting functions were developed by K. A. Peterson *et al.*,[126] F. Jensen,[127] and J. M. L. Martin[128] for specific basis sets, and/or specific systems. Even though the previous expressions were shown to be suitable for both the total E_{DFT} and the E_{HF} component,[111] the missing functions with angular momentum higher than i-functions suggests that the correlation energy itself should converge with an inverse power dependence.[26;129;130] As a consequence, many extrapolation procedures were developed as alternatives to eq. 2.41, and eq. 2.42 for extrapolation of the correlation energy.

In 1962, C. Schwarz[131] proposed an extrapolation procedure for energies of atoms based on an inverse power series function.

$$f\left(L_{max}\right) = f_{CBS} + AL_{max}^{-3} + BL_{max}^{-5} + CL_{max}^{-7} + \cdots \tag{2.43}$$

While eq. 2.43 provides accurate results for atoms, further approximations have to be made for molecules containing different types of atoms.[111;132;133] Later work from C. Schwartz resulted in three-point inverse fitting functions,

$$f\left(L_{max}\right) = f_{CBS} + \frac{A}{\left(L_{max}+\frac{1}{2}\right)^4} + \frac{B}{\left(L_{max}+\frac{1}{2}\right)^6} \tag{2.44}$$

$$f\left(L_{max}\right) = f_{CBS} + \frac{A}{\left(L_{max}+\frac{1}{2}\right)^\alpha} \tag{2.45}$$

and in two-point inverse fitting functions.

$$f\left(L_{max}\right) = f_{CBS} + \frac{A}{\left(L_{max}+\frac{1}{2}\right)^4} \tag{2.46}$$

The offset factor of 1/2 is a heuristic compromise between values of 0 for hydrogen and 1 for first-, and second-row elements. The argumentation behind this comes from the fact that the maximum angular momentum functions for hydrogen is one less than for first-, and second-row elements.
T. Helgaker *et al.* suggested a two-point scheme which requires two basis sets – N, and M – to compute the CBS correlation energy.[134]

$$f_{CBS} = \frac{N^3 \times f_N - M^3 \times f_M}{N^3 - M^3} \tag{2.47}$$

In 2003, H. F. Schaeffer III *et al.*[135] suggested a separate extrapolation of the singlet (eq. 2.48) and triplet (eq. 2.49) correlation energies.

$$f^{OS}\left(L_{max}\right) = f_{CBS} + \frac{A}{\left(L_{max}+\frac{1}{2}\right)^3} \tag{2.48}$$

$$f^{SS}\left(L_{max}\right) = f_{CBS} + \frac{A}{\left(L_{max}+\frac{1}{2}\right)^5} \tag{2.49}$$

Herein, we report the performance of eq. 2.41 and 2.42 denoted as F[2.41], and F[2.41], respectively, on the total energies and on the interaction energies. For the correlation part, we compare eq. 2.44,

2.45, and 2.46, denoted as S[2.44], S[2.45], and S[2.46], respectively.

At the CBS limit, the interaction energy E^{int}_{CBS} reads

$$E^{int}_{CBS} = E^{A\cdots B}_{CBS} - E^{A}_{CBS} - E^{B}_{CBS} \tag{2.50}$$

where $E^{A\cdots B}_{CBS}$, E^{A}_{CBS}, and E^{B}_{CBS} are respectively the total energy of the complex $A\cdots B$, the monomer A, and the monomer B at the CBS limit. The interaction energies E^{int}_{CBS} obtained with different combinations of extrapolation schemes are summarized in table 2.12. Nonetheless, it is of interest to discuss the dependence of the total energy on the extrapolation fits. For this purpose, the values obtained for $E^{A\cdots B}_{CBS}$ at the PBE, DSD-PBEPBE and MP2 level are displayed in Tables 2.9, 2.10, and 2.11, respectively, and further discussed in the next paragraphs.

Due to the square root in the F[2.42] extrapolation scheme, the $E^{A\cdots B}_{CBS}$ is systematically lower than the corresponding value obtained with the F[2.41]. On average, the difference in energy between the two schemes is of 1.207 mE_h, 0.723 mE_h, and 1.101 mE_h for PBE, DSD-PBEPBE, and MP2, respectively, with $\Delta E_{max} = 2.583$ mE_h (1.6 kcal/mol) for HCOOH$\cdots$HCOOH at the PBE level. By looking at the average R^2 value, the performance of the extrapolation functions on the five-point sequences can be appreciated. Despite the fact that they are both very similar ($R^2 = 0.9995$ for F[2.41], and $R^2 = 0.9990$ for F[2.42]), the F[2.41] has been shown to be a better fit in many cases,[125;134;136–138] and therefore, it is used to extrapolate the reference E_{CBS}/PBE and E_{CBS}/HF (bolt values in table 2.12).

Since the convergence of the correlation energy with respect to the basis set size is much slower than the convergence of the HF/DFT components, the dependence of $E^{A\cdots B}_{CBS}$/E(2) on the fit is more pronounced. A close look at table 2.10 and 2.11 reveals that, indeed, the average maximal difference $\Delta\bar{E}_{max}$, computed as

$$\Delta\bar{E}_{max} = \frac{1}{14}\sum_{i=1}^{14} \mathrm{MAX}_i \left\{ \left|[2.44] - [2.46]\right|, \left|2.44] - [2.45]\right|, \left|2.46] - [2.45]\right| \right\} \tag{2.51}$$

is much larger: $\Delta\bar{E}_{max} = 4.551$ mE_h for DSD-PBEPBE, and $\Delta\bar{E}_{max} = 8.231$ for MP2, with $\Delta E_{max} = 19.553$ mE_h (12.3 kcal/mol) for HCOOH$\cdots$HCOOH at the MP2 level. Inspection of the α-coefficient used in S[2.45] shows that a fixed coefficient, such as $\alpha = 4$ in eq. 2.46, is not a viable approach. Nevertheless, the near constant value of α over the 33 systems ($\bar{\alpha} = 3.317$ for DSD-PBEPBE, and $\bar{\alpha} = 3.312$ for MP2) indicates that S[2.45] captures some of the physics behind the data. However, even though R^2 of S[2.44] is closer to 1 than in the case of S[2.45], the fixed coefficients limit the scope of application to specific data, and, as a consequence, S[2.45] is used to extrapolate the reference E_{CBS}/E(2) for DSD-PBEPBE and MP2 (bolt values in table 2.12).

While absolute energies are the ultimate test for the various extrapolation schemes, relative energies are the main focus in most applications. Despite the observation of divergences in extrapolated energies, error cancellation reduces considerably the dependence of the interaction energy on the extrapolation scheme: E^{int}_{CBS} are computed within less than 200 μE_h.

Systems	$E_{CBS}^{A\cdots B}$/PBE in E_h			
	F[2.41]	R^2	F[2.42]	R^2
RG21 data set				
$He\cdots He$	-5.786	0.9994	-5.786	0.9996
$Ar\cdots Ar\cdots Ar$	-1582.035	0.9952	-1582.037	0.9940
CT7/04 data set				
$NH_3\cdots ClF$	-616.310	0.9992	-616.311	0.9998
$NH_3\cdots F_2$	-255.962	1.0000	-255.964	0.9998
HB9 data set				
$HCOOH\cdots HCOOH$	-379.350	0.9999	-379.353	0.9995
$NH_3\cdots NH_3$	-113.039	1.0000	-113.040	0.9999
ADIM4 data set				
$CH_4\cdots CH_4$	-80.936	0.9998	-80.937	1.0000
DI9 data set				
$CH_3SH\cdots HCl$	-899.165	0.9962	-899.167	0.9974
$CH_3Cl\cdots CH_2O$	-614.340	0.9950	-614.341	0.9980
DISP4 data set				
anti	-158.288	1.0000	-158.289	0.9980
gauche	-158.287	1.0000	-158.288	0.9980
syn	-158.280	1.0000	-158.281	0.9980
PPS11 data set				
$CH_2\cdots CH_2$	-154.519	0.9998	-154.520	1.0000
$CH_2\cdots CH_2$	-154.514	0.9998	-154.515	1.0000
overall				
Average R^2	–	0.9989	–	0.9987
Difference ΔE	$\Delta\bar{E} = 1.207$, $\Delta E_{min} = 0.019$, $\Delta E_{max} = 2.583$			

Table 2.9: $E_{CBS}^{A\cdots B}$ at the PBE level of theory. The total energy is extrapolated with the Feller schemes of eq. 2.41, and eq. 2.42. $E_{CBS}^{A\cdots B}$ are in E_h, and ΔE in mE_h.

Systems	$E_{CBS}^{A\cdots B}$/PBE in E_h				$E_{CBS}^{A\cdots B}$/E(2) in mE_h						
	F[2.41]	R^2	F[2.42]	R^2	S[2.44]	R^2	S[2.45]	α	R^2	S[2.46]	R^2
RG21 data set											
$He\cdots He$	-5.758	0.9994	-5.758	0.9996	-42.425	0.9990	-42.549	3.370	0.9984	-42.130	0.9948
$Ar\cdots Ar\cdots Ar$	-1581.344	0.9999	-1581.345	0.9996	-354.620	0.9940	-357.566	3.001	0.9909	-348.758	0.9822
CT7/04 data set											
$NH_3\cdots ClF$	-615.754	1.0000	-615.755	0.9996	-387.467	0.9985	-389.635	3.090	0.9976	-383.128	0.9903
$NH_3\cdots F_2$	-255.474	1.0000	-255.476	0.9997	-421.470	0.9990	-424.108	3.031	0.9981	-416.612	0.9900
HB9 data set											
$HCOOH\cdots HCOOH$	-378.546	0.9999	-378.548	0.9994	-686.797	0.9987	-690.342	3.086	0.9977	-679.784	0.9904
$NH_3\cdots NH_3$	-112.757	1.0000	-112.757	0.9997	-259.311	0.9989	-260.107	3.293	0.9984	-257.430	0.9943
ADIM4 data set											
$CH_4\cdots CH_4$	-80.704	0.9999	-80.704	1.0000	-222.739	0.9990	-223.146	3.460	0.9987	-221.594	0.9965
DI9 data set											
$CH_3SH\cdots HCl$	-898.586	1.0000	-898.587	0.9999	-324.032	0.9983	-325.390	3.209	0.9975	-320.927	0.9923
$CH_3Cl\cdots CH_2O$	-613.747	1.0000	-613.748	0.9997	-428.822	0.9985	-430.737	3.163	0.9976	-424.700	0.9917
DISP4 data set											
anti	-157.861	1.0000	-157.861	0.9970	-394.130	0.9989	-394.903	3.434	0.9985	-391.989	0.9961
gauche	-157.860	1.0000	-157.860	0.9970	-394.505	0.9989	-395.271	3.438	0.9985	-392.378	0.9962
syn	-157.851	1.0000	-157.852	0.9970	-394.875	0.9989	-395.637	3.440	0.9985	-392.754	0.9962
PPS11 data set											
$CH_2\cdots CH_2$	-154.126	1.0000	-154.127	0.9999	-343.347	0.9985	-344.261	3.312	0.9978	-341.027	0.9941
$CH_2\cdots CH_2$	-154.121	1.0000	-154.121	0.9999	-344.545	0.9985	-345.477	3.306	0.9978	-342.195	0.9940
overall											
Average R^2	–	0.9999	–	0.9991	–	0.9984	–	3.317	0.9976	–	0.9928
Difference ΔE	$\Delta\bar{E} = 0.723$, $\Delta E_{min} = 0.018$, $\Delta E_{max} = 1.794$				$\Delta\bar{E}_{min} = 1.432$, $\Delta\bar{E}_{max} = 4.551$, $\Delta E_{min} = 0.124$, $\Delta E_{max} = 10.558$						

Table 2.10: $E_{CBS}^{A\cdots B}$ at the DSD-PBEPBE level of theory. The total DFT energy is extrapolated with the Feller schemes of eq. 2.41, and eq. 2.42 and the spin-component-scaled MP2 correlation energy with the Schwartz schemes of eq. 2.44, eq. 2.45, and eq. 2.46. $E_{CBS}^{A\cdots B}$/PBE are in E_h while $E_{CBS}^{A\cdots B}$/E(2), and ΔE are in mE_h.

Systems	$E_{CBS}^{A\cdots B}$/SCF in E_h				$E_{CBS}^{A\cdots B}$/E(2) in mE_h						
	F[2.41]	R^2	F[2.42]	R^2	S[2.44]	R^2	S[2.45]	α	R^2	S[2.46]	R^2
RG21 data set											
$He\cdots He$	-5.723	0.9994	-5.723	0.9995	-73.769	0.9990	-74.015	3.335	0.9983	-73.204	0.9942
$Ar\cdots Ar\cdots Ar$	-1580.452	0.9996	-1580.453	0.9980	-729.458	0.9948	-734.488	3.075	0.9923	-718.929	0.9851
CT7/04 data set											
$NH_3\cdots ClF$	-615.148	0.9999	-615.149	0.9995	-789.891	0.9987	-793.720	3.156	0.9979	-781.817	0.9917
$NH_3\cdots F_2$	-254.996	1.0000	-254.998	0.9997	-863.048	0.9991	-867.714	3.107	0.9983	-853.935	0.9915
HB9 data set											
$HCOOH\cdots HCOOH$	-377.739	0.9999	-377.740	0.9995	-1417.003	0.9987	-1423.323	3.153	0.9979	-1403.790	0.9917
$NH_3\cdots NH_3$	-112.452	0.9989	-112.452	0.9988	-522.844	0.9987	-524.082	3.395	0.9983	-519.674	0.9955
ADIM4 data set											
$CH_4\cdots CH_4$	-80.433	0.9999	-80.434	1.0000	-433.420	0.9991	-434.120	3.502	0.9988	-431.355	0.9970
DI9 data set											
$CH_3SH\cdots HCl$	-897.881	1.0000	-897.881	0.9997	-645.374	0.9985	-647.725	3.264	0.9978	-639.757	0.9934
$CH_3Cl\cdots CH_2O$	-613.079	1.0000	613.080	0.9997	-863.280	0.9986	-866.668	3.218	0.9979	-855.643	0.9928
DISP4 data set											
anti	-157.368	1.0000	-157.368	0.9980	-784.043	0.9990	-785.373	3.482	0.9986	-780.158	0.9967
gauche	-157.366	1.0000	-157.367	0.9980	-784.928	0.9990	-786.244	3.486	0.9986	-781.070	0.9967
syn	-157.358	1.0000	-157.358	0.9980	-785.302	0.9990	-786.609	3.488	0.9986	-781.454	0.9967
PPS11 data set											
$CH_2\cdots CH_2$	-153.711	1.0000	-153.712	0.9999	-681.153	0.9985	-682.768	3.358	0.9980	-676.850	0.9948
$CH_2\cdots CH_2$	-153.705	1.0000	-153.705	0.9999	-683.656	0.9985	-685.306	3.352	0.9980	-679.290	0.9947
overall											
Average R^2	–	0.9998	–	0.9992	–	0.9985	–	3.312	0.9978	–	0.9938
Difference ΔE	$\Delta\bar{E} = 1.101$, $\Delta E_{min} = 0.018$, $\Delta E_{max} = 1.611$				$\Delta\bar{E}_{min} = 2.499$, $\Delta\bar{E}_{max} = 8.231$, $\Delta E_{min} = 0.246$, $\Delta E_{max} = 19.533$						

Table 2.11: $E_{CBS}^{A\cdots B}$ at the MP2 level of theory. The total HF energy is extrapolated with the Feller schemes of eq. 2.41, and eq. 2.42 and the MP2 correlation energy with the Schwartz schemes of eq. 2.44, eq. 2.45, and eq. 2.46. $E_{CBS}^{A\cdots B}$/SCF are in E_h while $E_{CBS}^{A\cdots B}$/E(2), and ΔE are in mE_h.

Systems	E^{int}_{CBS}/PBE		E^{int}_{CBS}/DSD-PBEPBE						E^{int}_{CBS}/MP2					
	F[2.41]	F[2.42]	F[2.41]			F[2.42]			F[2.41]			F[2.42]		
			S[2.44]	S[2.45]	S[2.46]	S[2.44]	S[2.45]	S[2.46]	S[2.44]	S[2.45]	S[2.46]	S[2.44]	S[2.45]	S[2.46]
RG21 data set														
$He \cdots He$	-0.06	-0.06	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
$Ar \cdots Ar \cdots Ar$	-0.29	-0.29	-0.77	-0.76	-0.78	-0.76	-0.75	-0.77	-0.97	-0.97	-0.98	-0.97	-0.96	-0.97
CT7/04 data set														
$NH_3 \cdots ClF$	-16.74	-16.73	-12.05	-11.99	-12.08	-12.05	-11.99	-12.08	-12.00	-11.88	-12.05	-12.00	-11.89	-12.05
$NH_3 \cdots F_2$	-4.97	-4.97	-1.86	-1.81	-1.88	-1.86	-1.81	-1.87	-1.84	-1.76	-1.86	-1.83	-1.75	-1.86
HB9 data set														
$HCOOH \cdots HCOOH$	-18.27	-18.26	-18.79	-18.69	-18.89	-18.77	-18.67	-18.87	-18.57	-18.39	-18.76	-18.55	-18.37	-18.74
$NH_3 \cdots NH_3$	-2.84	-2.84	-3.28	-3.27	-3.30	-3.28	-3.27	-3.30	-2.34	-2.23	-2.54	-2.30	-2.20	-2.50
ADIM4 data set														
$CH_4 \cdots CH_4$	-0.10	-0.10	-0.52	-0.51	-0.52	-0.52	-0.51	-0.52	-0.50	-0.49	-0.51	-0.50	-0.49	-0.51
DI9 data set														
$CH_3SH \cdots HCl$	-5.41	-5.41	-5.11	-5.06	-5.16	-5.11	-5.06	-5.16	-5.49	-5.40	-5.59	-5.49	-5.40	-5.59
$CH_3Cl \cdots CH_2O$	-0.51	-0.50	-1.05	-1.03	-1.08	-1.05	-1.02	-1.08	-1.26	-1.21	-1.31	-1.26	-1.21	-1.31
DISP4 data set														
anti→gauche	0.82	0.82	0.53	0.53	0.52	0.53	0.53	0.52	0.58	0.59	0.57	0.58	0.59	0.57
anti→syn	5.53	5.53	5.36	5.37	5.35	5.36	5.37	5.35	5.63	5.65	5.61	5.63	5.65	5.61
PPS11 data set														
$CH_2 \cdots CH_2$	-1.24	-1.24	-1.64	-1.62	-1.65	-1.63	-1.62	-1.65	-1.66	-1.64	-1.69	-1.65	-1.63	-1.69
$CH_2 \cdots CH_2$	1.86	1.86	1.10	1.10	1.10	1.10	1.10	1.10	0.65	0.65	0.65	0.65	0.65	0.66

Table 2.12: PBE, DSD-PBEPBE, and MP2 interacting energies (in kcal/mol) of the strongest and weakest interacting complexes at the complete basis set limit with different combination of extrapolation schemes. Values on a grey background are used for further statistical analysis of the basis sets.

The assessment of the basis set performance on the interaction energies of the strongest and weakest interacting complexes is carried out by means of statistical analysis, through the normal distribution around the CBS limit (see Fig. 2.10, and Table 2.13):

$$\Theta_{MAE,\,\sigma}(X) = \frac{1}{\sigma\sqrt{2\Pi}} \exp\left(-\frac{1}{2}\frac{(X-MAE)^2}{\sigma^2}\right) \tag{2.52}$$

where the mean absolute error MAE reads

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\left(E^{int}_{CBS,i} - E^{int}_{M,i}\right)^2} \tag{2.53}$$

and the standard deviation σ

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\sqrt{\left(E^{int}_{CBS,i} - E^{int}_{M,i}\right)^2} - MAE\right)^2} \tag{2.54}$$

with N the number of complexes forming the set (i.e. $N = 13$ interaction energies), $E^{int}_{CBS,i}$ the CBS interaction energy of system i, and $E^{int}_{M,i}$ the adsorption energy of system i computed at the M basis set level.

Table 2.13 shows also the raw average error of the M basis set to recover both the absolute HF/DFT (eq. 2.55), and correlation energies (eq. 2.56) of the 33 systems, which are computed as

$$\Delta\bar{E}_{HF/DFT} = \frac{1}{N}\left(E^{abs}_{HF/DFT,M} - E^{abs}_{HF/DFT,CBS}\right) \tag{2.55}$$

and

$$\Delta\bar{E}_{E(2)} = \frac{1}{N}\left(E^{abs}_{E(2),M} - E^{abs}_{E(2),CBS}\right) \tag{2.56}$$

with $N = 33$ (11 dimers and 22 monomers).

Inspecting the MAE (Table 2.13) reveals a rapid convergence of the interaction energies towards the CBS. However, $\Delta\bar{E}_{HF/DFT,M}$ shows that a basis set of a least quadruple-zeta quality is required to be by less than 1.6 mE_h (≈1 kcal/mol) away from the CBS. Nevertheless, even though the average is bellow 1.6 mE_h. The maximum gap between the aug-cc-pCVQZ basis and the CBS is 4.72 mE_h. At the sextuple level, the $\Delta\bar{E}_{HF/DFT}$ drops down to 0.21 mE_h (≈0.1 kcal/mol), with a maximal difference of $\Delta E_{HF/DFT} = 0.79$ mE_h. The convergence for dimer energies is much slower and even at the sextuple level, the total energies differ by 0.37 mE_h (≈0.2 kcal/mol) from the CBS value. This difference in convergence rate is reflected in both the σ and the MAE parameters of the normal distribution $\Theta_{MAE,\,\sigma}(X)$. Since the latrer is computed on the interaction energies around the E^{int}_{CBS} of (mainly) dimers, a good estimation of the error cancellation can be given by the quantity Q defined in eq. 2.57.

$$\sigma,\ MAE\ \propto Q = \Delta E^{dimers}_{HF/DFT,M} - 2\times\Delta E^{monomers}_{HF/DFT,M} \tag{2.57}$$

Consequently, because Q at the aug-cc-pV5Z level is very similar to Q at the aug-cc-pV6Z (*ca.* 10 μE_h), $\Theta_{MAE,\,\sigma}(X)$ differs by only a few cal/mol (see table 2.13).

The convergence of the correlation energy is, as expected, much slower than the HF/DFT components: even with the large aug-cc-pV6Z basis set, the gap with the CBS limit is of 4.48 mE_h for the monomers and 8.25 mE_h for the dimers. Surprisingly, the Def2-SVP and Def2-SVPD basis perform better on the E(2) component than on the HF/DFT part, which is most likely accidental.

In either cases, a systematic reduction of both the gap with the CBS limit, and the parameters of the normal distribution is obtained with the increase of diffuse and polarised functions. Indeed, as shown in figure 2.8, the field experienced by monomer A from monomer B results in a tendency of the electrons to shift away from the nucleus. Typically, such electron displacement is described by basis sets augmented with higher L quantum number functions than the original ones: the polarization functions. Moreover, including diffuse functions to the basis sets is essential for an accurate representation of the outer region of the charge density cloud.

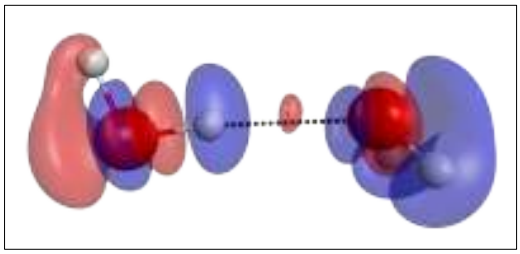


Figure 2.8: Representation of the electronic rearrangement upon formation of the water dimer. Contour of 0.001 electron/Å^3 is shown. In red the electron accumulation, in blue the electron depletion.

Addition of [1s1p] steep (tight) diffuse correlating functions to the aug-cc-pVDZ (see 1-Row description in table 2.8) reduces the gap aug-cc-pVDZ/aug-cc-pCVDZ by up to 5.04 mE_h for the correlation component of the dimers.

A similar observation is made when aug-cc-pVTZ is augmented with [2s2p1d] steep functions: the gap is reduced to 4.56 mE_h at most. On the interaction energies, the MAE reflects an improvement by one order of magnitude for the HF/DFT component and is reduced to 0.1 kcal/mol for the correlation part.

At the quadruple-ζ level, the gap is only 2.57 mE_h when [3s3p2d1f] functions are added to aug-cc-pVQZ, and the MAE is further improved and reached 0.012 kcal/mol at its best. As L_{max} increases, the aug-cc-pVnZ/aug-cc-pCVnZ gap is expected to decrease, and to vanish at the CBS limit. Not only the $\Delta\bar{E}^{abs}_{HF/DFT,M}$ and $\Delta\bar{E}^{abs}_{E(2),M}$ are improved by inclusion of steep functions, but $\Theta_{MAE,\,\sigma}(X)$ gets narrower and moves towards the CBS limit (see table 2.13, and fig. 2.10). However, because double-ζ basis is too small for a decent description of correlation effects, the improvement of tight diffuse functions is rather heterogeneous, yielding meaningless DSD-PBEPBE and MP2 normal distributions.

From Def2-SVP to Def2-SVPD, an additional set of correlating functions of each angular momentum is included, which results in considerable improvement: the gap with the CBS value is reduced by 21.27 mE_h in the best scenario. However, it is important to mention that with energies by up to 290 mE_h (*ca.* 180 kcal/mol) away from the CBS, the split-valence set is not suitable for calculations on correlated systems. Even the MAE, which benefits from error cancellation, is equal to *ca.* 1 kcal/mol.

The absolute energies are greatly improved when moving to the triple-ζ Def2-TZVP basis. Further improvements are perceived by addition of [1s1p1d] functions to either the Def2-TZVP or the Def2-TZVPP set, to form respectively the Def2-TZVPD and Def2-TZVPPD sets. $\Delta\bar{E}_{E(2)}$ is reduced by up to 3.86 mE_h. Concerning the error on the interaction energies, Def2-TZVPPD gives MAE between 0.081 kcal/mol and 0.174 kcal/mol.

At the quadruple-ζ level the absolute energies get closer to the CBS limit, and the addition of polarized

and/or diffuse functions only improves the total energies by 0.57 mE_h at most. It is important to stress out that at the Def2-QZVP level, no polarization functions are added to the 1-Row and 2-Row elements (see table 2.8). Hence, the pair Def2-QZVP/Def2-QZVPP and Def2-QZVPD/Def2-QZVPPD present the same σ and MAE. The best estimate of the interaction energy is provided by Def2-QZVPD with MAE between 0.011 and 0.165 kcal/mol. From Def2-QZVP to Def2-QZVPD and to def-QZVPPD, or from Def2-QZVPP to Def2-QZVPPD, [1s1p1d] are added, and this recovers the total energies by similar amounts.

From the calculations presented in this section, it is shown that for high accuracy interaction energies, basis sets containing at least g functions for the 1-Row and 2-Row are required. Among the six quadruple-ζ basis sets included in this study, the absolute energies of the HF and DFT components are best evaluated with the Def2-QZVPD basis, while the aug-cc-pCQZ provides the best results for the correlation component. In either case, the $\Delta\bar{E}$ challenges the aug-cc-pV5Z results. However, since the main focus is on interaction energies, the normal distribution around the E_{CBS}^{int} is heavier in the assessment process. With an MAE between 0.011 and 0.165 kcal/mol, Def2-QZVPD outperforms aug-cc-pVQZ by a few cal/mol. A last element to consider in the assessment is the computational effort. Figure 2.9 displays the computational cost in function of the interaction energies MAE, and this sheds light on the performance of the Def2-QZVPD basis set to provide the best ratio between the maximum amount of correlation energy recovered and *low* computational costs. As a consequence, the Def2-QZVPD basis set will be used to assess the performance of a wide range of N-electron models on weakly interacting complexes.

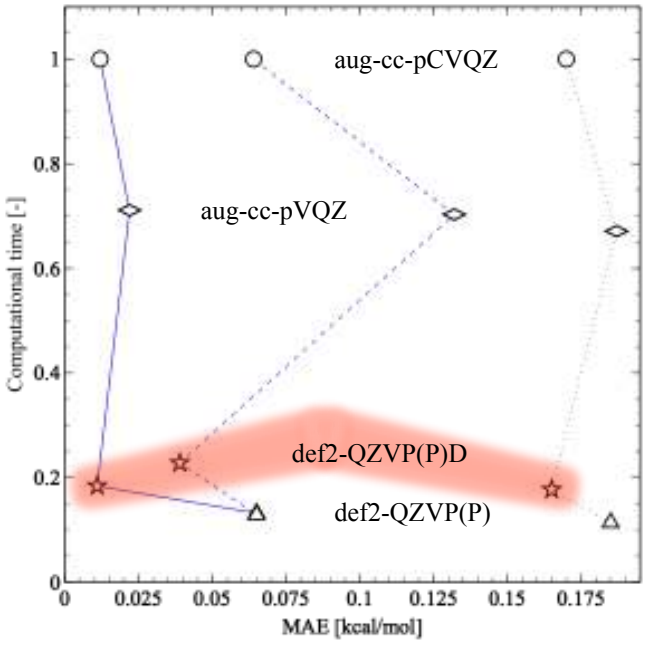


Figure 2.9: Computational time of the quadruple-ζ basis: aug-cc-pCVQZ ($\circ$), aug-cc-pVQZ ($\diamond$), Def2-QZVP and Def2-QZVPP ($\star$), Def2-QZVPD and Def2-QZVPPD ($\triangle$) in function of the MAE. To guide the eye, the solid line shows the PBE results, the dashed line shows the DSD-PBEPBE results, and the doted line shows the MP2 results.

Basis set	$\Delta\bar{E}^{abs}_{HF/DFT,M}$ in mE_h		$\Delta\bar{E}^{abs}_{E(2),M}$ in mE_h		Normal distribution in kcal/mol					
	dimers	monomers	dimers	monomers	PBE		DSD-PBEPBE		MP2	
					σ	MAE	σ	MAE	σ	MAE
Double-zeta										
aug-cc-pVDZ	62.31	36.16	137.10	78.15	0.348	0.348	0.316	0.362	0.340	0.380
aug-cc-pCVDZ	59.99	34.66	132.06	75.25	0.244	0.300	0.461	0.464	0.867	0.598
Def2-SVP	290.10	149.16	162.90	91.12	1.679	1.358	1.161	0.888	0.925	0.682
Def2-SVPD	272.97	140.55	141.64	79.75	0.652	0.838	0.636	0.881	0.598	0.956
Triple-zeta										
aug-cc-pVTZ	15.87	9.18	49.07	27.36	0.047	0.049	0.121	0.129	0.291	0.234
aug-cc-pCVTZ	13.28	7.68	44.51	25.02	0.047	0.041	0.106	0.106	0.272	0.202
Def2-TZVP	20.72	10.63	70.81	40.72	0.281	0.274	0.155	0.128	0.325	0.328
Def2-TZVPD	19.13	9.78	66.95	38.70	0.098	0.105	0.083	0.118	0.270	0.212
Def2-TZVPP	18.48	9.17	60.07	33.38	0.253	0.243	0.153	0.142	0.371	0.275
Def2-TZVPPD	17.15	8.48	57.09	31.94	0.078	0.081	0.054	0.081	0.246	0.174
Quadruple-zeta										
aug-cc-pVQZ	4.37	2.46	17.37	9.66	0.033	0.022	0.189	0.132	0.270	0.187
aug-cc-pCVQZ	2.55	1.52	14.80	8.39	0.014	0.012	0.072	0.064	0.257	0.170
Def2-QZVP	1.81	1.15	22.55	12.53	0.072	0.065	0.082	0.065	0.269	0.185
Def2-QZVPD	1.58	1.01	21.98	12.26	0.010	0.011	0.052	0.039	0.241	0.165
Def2-QZVPP	1.81	1.15	22.55	12.53	0.072	0.065	0.082	0.065	0.269	0.185
Def2-QZVPPD	1.58	1.01	21.98	12.26	0.010	0.011	0.052	0.039	0.241	0.165
Quintuple-zeta										
aug-cc-pV5Z	1.03	0.56	11.05	6.01	0.002	0.004	0.043	0.040	0.225	0.132
Sextuple-zeta										
aug-cc-pV6Z	0.37	0.21	8.25	4.48	0.005	0.007	0.031	0.025	0.242	0.115

Table 2.13: The average difference in absolute energies, $\Delta\bar{E}^{abs}_{HF/DFT}$ and $\Delta\bar{E}^{abs}_{E(2)}$, over the 33 systems (11 complexes, and the 22 isolated molecules) from the CBS energies are in mE_h. The results of the normal distribution are in kcal/mol and they have been computed around the 13 interaction energies at the CBS limit.

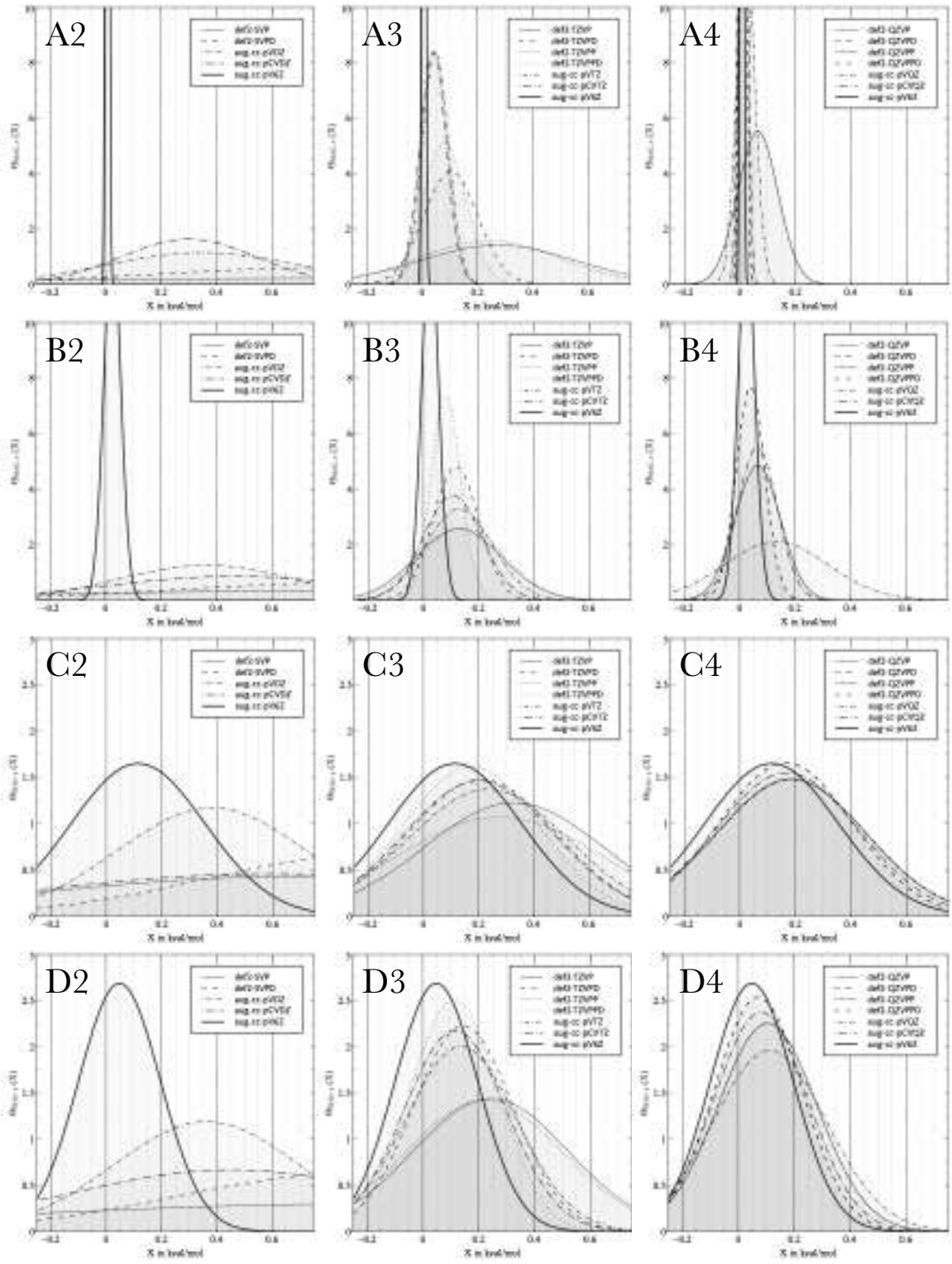


Figure 2.10: Normal distribution of the interaction energies around the CBS value for PBE (A), DSD-PBEPBE (B), MP2 (C), and overall (D) with the Ahlrichs and Dunning basis sets. For the sake of clarity, the results have been separated into the corresponding double-zeta (2), triple-zeta (3), and quadruple-zeta (4) basis sets. In all the figures, the aug-cc-pV6Z result is displayed (solid black line) to appreciate the improvement of the results by increasing the number of gaussian basis functions. It is important to note that because the pairs Def2-QZVP/Def2-QZVPP and defQZVPD/Def2-QZVPPD have same σ and MAE values, the same line display is used.

B. Auxiliary basis set

To assess the auxiliary basis set, DSD-PBEPBE/Def2-QZVPD has been applied on the reduce set (Fig. 2.7). The cardinal number relation (see eq. 2.41) has been used to compute the $E_{corr.}^{CBS}$.

Fixing `SUBROUTINE RIMP2CSTRM` in GAMESS

On the way to reaching the CBS limit, auxiliary basis sets (ABS) involving spherical H shells and higher angular momentum lead to final energies too large to be expressed in a standard GAMESS output file, as shown in table 2.14

ABS	L_{max}^{H}	L_{max}^{C}	$E_c^{(2)}$
aug-cc-pVDZ	D	F	`E(2)= -0.2089440218`
aug-cc-pVTZ	F	G	`E(2)= -0.2111151132`
aug-cc-pVQZ	G	H	`E(2)= **********`
aug-cc-pV5Z	H	I	`E(2)= **********`
Reference: MP2/Def2-QZVPD			`E(2)= -0.2097717833`

Table 2.14: RI-MP2 correlation energy $E_c^{(2)}$ of CH_4 at different ABS level. L_{max}^{H} and L_{max}^{C} are the highest angular momentum for atom C and H, respectively. Values are in Ha, and were obtained before fixing `SUBROUTINE RIMP2CSTRM`.

A close look ar the `SUBROUTINE RIMP2CSTRM` revealed a few errors in the implementation of the RI approximation. In particular, this required working out various transformation matrices from Cartesian to spherical coordinates. This starts with adding the missing common block for H and I shell. As shown below, the new common block called `SPHEHI` was added after the existing `SPHERI` block.

```
COMMON /SPHEHI/  HSHELL(21,21), AISHELL(28,28),
*                HIHELL(21,21), AIIHELL(28,28)

COMMON /SPHERI/  PSHELL(3,3),DSHELL(6,6),
*                FSHELL(10,10),GSHELL(15,15),
*                PIHELL(3,3),DIHELL(6,6),
*                FIHELL(10,10),GIHELL(15,15)
```

This addition lead to the addition of two extra loops to gather data for the missing H and I shells in various subroutines.

```
[...]
ELSE IF(IT .EQ. 5) THEN
  MINIS = 26
  MAXIS = 36
  DO I = 1, 21
    DO J = 1, 11
      TR(I,J) = HSHELL(I,J)
    END DO
  END DO
```

```
ELSE IF(IT .EQ. 6) THEN
  MINIS = 37
  MAXIS = 49
  DO I = 1, 28
    DO J = 1, 13
      TR(I,J) = AISHELL(I,J)
    END DO
  END DO
[...]
```

The values of the new common block are initialized in `SUBROUTINE GETHROT` and `SUBROUTINE GETIROT`, called by `SUBROUTINE SPHSET` in `symslc.src`. Last but not least, the storage allocated to `TR` needed to be carefully checked in order to make sure that the first argument in the subroutine is big enough for H and I shells.

Finally, as an ultimate validation, RI-MP2/Def2-QZVPDD with aug-cc-pV5Z as ABS on CH_4 gives `E(2)= -0.2112361754` which proves the successful fix of the RI-approximation (see table 2.14 for a comparison).

Performance of the auxiliary basis set

The performance of the auxiliary basis set (ABS) is established on the deviation of the correlation energy within the RI-approximation from the standard four-index two electron integrals (see eq. 2.40, and section III). Table 2.15 displays the performance of the ten tested auxiliary basis set (ABS) on the reduced data set (Fig. 2.7), at the MP2 level. We introduce a CBS-*like* limit referred to as CABS (complete auxiliary basis set). As was the case for the CBS in the previous section, the correlation energy component was extrapolated to the CABS following C. Schwartz.

$$f\left(L_{max}\right) = f_{CBS} + \frac{A}{\left(L_{max} + \frac{1}{2}\right)^{\alpha}} \tag{2.58}$$

Table 2.15 shows the rapid convergence of the ABS to the CABS. As a matter of fact, most of the triple-ζ ABS have already reached CABS accuracy. Nonetheless, for the sake of consistency a quadruple-ζ ABS should be used. In this regards, only cc-pVQZ and Def2-QZVP are available. For the same reasons mentioned in section VA., the computational effort favors Def2-QZVP, which will be used to extensively study the cost and accuracy of the RI-approximation in section IX.

Auxiliary basis set	$He\cdots He$	$Ar\cdots Ar$	$NH_3\cdots ClF$	$NH_3\cdots F_2$	$HCOOH\cdots HCOOH$	$NH_3\cdots NH_3$	$CH_4\cdots CH_4$	$CH_3SH\cdots HCl$	$CH_3Cl\cdots CH_2O$	anti $\rightarrow$ gauche	anti $\rightarrow$ syn	$CH_2\cdots CH_2$	$CH_2\cdots CH_2$
Double-zeta													
cc-pVDZ	-0.01	-0.82	-11.78	-1.77	-19.33	-3.17	-0.51	-5.51	-1.29	0.59	5.04	-1.79	0.73
aug-cc-pVDZ	-0.01	-0.81	-11.76	-1.76	-19.33	-3.17	-0.54	-5.46	-1.29	0.58	5.02	-1.76	0.72
Def2-SVP	-0.01	-0.78	-11.77	-1.77	-19.27	-3.18	-0.53	-5.52	-1.27	0.59	5.04	-1.73	0.79
Def2-SVPD	-0.01	-0.77	-11.68	-1.75	-19.20	-3.18	-0.54	-5.42	-1.27	0.58	5.03	-1.74	0.75
Triple-zeta													
cc-pVTZ	-0.01	-0.78	-11.65	-1.76	-19.24	-3.15	-0.52	-5.42	-1.30	0.55	5.01	-1.78	0.74
aug-cc-pVTZ	-0.01	-0.78	-11.61	-1.75	-19.23	-3.14	-0.52	-5.40	-1.29	0.55	5.01	-1.77	0.74
Def2-TZVP	-0.01	-0.78	-11.67	-1.76	-19.33	-3.17	-0.52	-5.46	-1.30	0.54	5.01	-1.79	0.73
Def2-TZVPD	-0.01	-0.78	-11.65	-1.76	-19.30	-3.17	-0.53	-5.45	-1.30	0.53	5.01	-1.78	0.74
Def2-TZVPP	-0.01	-0.78	-11.65	-1.76	-19.24	-3.15	-0.52	-5.42	-1.30	0.54	5.01	-1.78	0.74
Def2-TZVPPD	-0.01	-0.77	-11.63	-1.75	-19.23	-3.15	-0.52	-5.41	-1.29	0.54	5.01	-1.77	0.74
Quadruple-zeta													
cc-pVQZ	-0.01	-0.77	-11.61	-1.75	-19.23	-3.14	-0.51	-5.40	-1.30	0.55	5.01	-1.78	0.74
Def2-QZVP	-0.01	-0.77	-11.61	-1.75	-19.23	-3.14	-0.51	-5.40	-1.30	0.55	5.01	-1.78	0.74
Extrapolated deviation, CBS limit													
CBS	-0.01	-0.77	-11.59	-1.75	-19.23	-3.13	-0.51	-5.40	-1.30	0.55	5.01	-1.78	0.74
Reference values: MP2/Def2-QZVPD													
E_{ref}^{int}	-0.01	-0.97	-11.88	-1.76	-18.39	-2.23	-0.49	-5.40	-1.21	0.59	5.65	-1.64	0.65
$\left\lvert E_{ref}^{int} - E_{best}^{int} \right\rvert$	0.00	0.15	0.10	0.00	0.81	0.91	0.02	0.00	0.06	0.00	0.61	0.09	0.07

Table 2.15: Deviation of RI-MP2/Def2-QZVPD at various auxiliary basis set level from the reference MP2/Def2-QZVPD. Energies are in kcal/mol.

VI Method development and implementation

From the initial GGA implementation scheme shown in Fig. 2.11, the contributions detailed in the following sections lead to the final implementation depicted in Fig. 2.12.

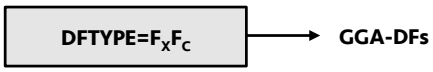


Figure 2.11: Initial implementation scheme of the GGAs.

As already mentioned in Introduction, these contributions consist in implementing new methodologies and enhancing existing quantum chemical methods in GAMESS. These contributions are briefly listed below.

- ▷ (RI-)DSD-DFTs: implementation of 300 dispersion-corrected spin-component-scaled double-hybrid DFTs and their corresponding cost-effective RI version (section VII and IX).
- ▷ (RI-)SCS-DFTs: implementation and optimization of the spin-component-scaled double-hybrid DFTs and their corresponding cost-effective RI version (section VIII and IX).
- ▷ MP2($erfc$): implementation of an attenuated MP2 in which the standard Coulomb operator is replaced by the complementary error function $erfc$ (section X).
- ▷ DH($erfc$) DFTs: implementation of double-hybrids with the attenuated version of MP2 (section XI).

These contributions required considerable changes in common blocks, keywords, branching, *etc.* yielding a drastic modification of the GGA implementation scheme, as illustrated in Fig. 2.12.

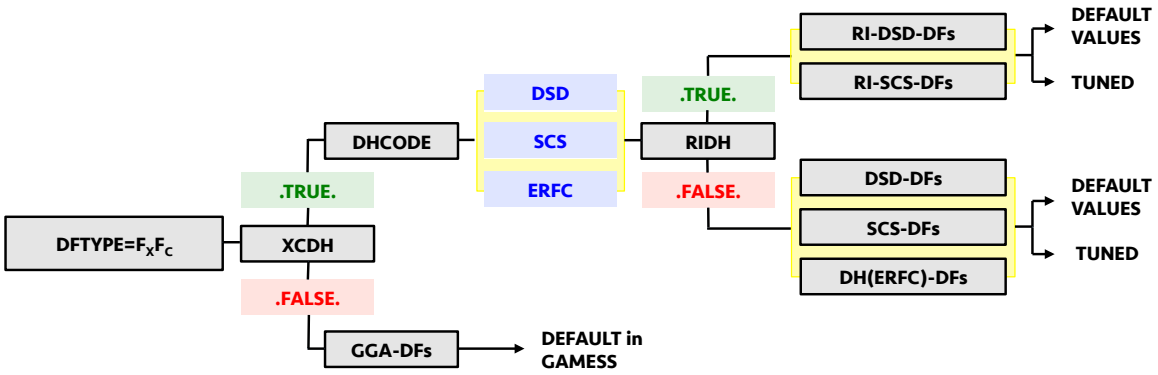


Figure 2.12: Simplified overall final implementation scheme.

In the following sections each contributions and their relative performance is considered separately. The reader has to bear in mind that the theoretical background introduced earlier, as well as the exhaustive basis set study considered in section V are greatly cross-referenced along the sections.

VII Implementation and performance of dispersion-corrected spin-component-scaled DFTs and post Hartree-Fock methods on non-bonded complexes

Among the variety of quantum chemical theories, Kohn-Sham density functional theory[8;9] (DFT) has established itself as the theoretical method of choice to account for electron correlation, notably because of its favorable computational cost over accuracy ratio. Although DFT theory is exact, approximations are made for the electron interactions. Thus, the success of DFT critically depends on the quality of the exchange-correlation functional E_{XC}. Despite the success of GGA functionals (such as PBE[71] and BLYP[66;67]) over the last 20 years, it is widely known that they do not provide an adequate description of non-local dispersion forces.[139;31] Recently, however, to overcome the lack of dispersion interactions in the GGA functionals, several methods have been developed, allowing the computational community to probe systems in which these interactions are crucial[140] (see section I).

A key aim of this section is to clarify the significance of electron correlation in the dimerization process of nonbonded complexes and to understand how this differs from one method to another. In what follows, we assess the performance of nine GGA DFs, one meta-GGA DF, two hybrid DFs, one meta-hybrid-GGA DF, two range-separated DFs, and ten double hybrid DFs. To assess the performance of the various DFT methods, the present study includes four wave function (WFT) methods: Hartree-Fock, Møller-Plesset second order perturbation theory (MP2), spin-component-scaled MP2, and Coupled-Cluster CCSD(T).

The seven data sets introduced in section IV are used to compare, and to assess the accuracy of various DFTs and WFTs.

In the next sections, we first outline the computational setup used, and present the implementation scheme of DSD-DFTs[12;13] into GAMESS[14]. It follows with the presentation of the results and analysis for the interaction energies of the HB9, CT7/04, DI9, ADIM5, IDISP4, PPS11, and RG21 data sets (introduced in section IV). In the last section, we summarize our results.

A. Implementation scheme into GAMESS

The implementation of the DSD-DFTs was carried out in this thesis work as a first approach to density functional implementations. The initial GGA implementation scheme show nin Fig. 2.12 of the previous section was modified to enable branching from GGA to DSD. The simplified implementation scheme is depicted in Fig. 2.13.

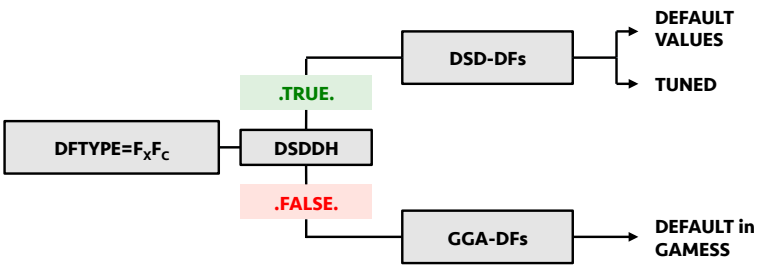


Figure 2.13: Implementation scheme of the DSD-DFs.

The new logical keyword `DSDDH` is used to branch standard GGA and DSD schemes. Aiming at a general implementation, all *ca.* 300 combinations of exchange-correlation GGA functionals (DFTYPE=F_XF_C)

in Fig. 2.13 can run either as a stand-alone GGA or as a DSD-DFT. The former is the default in GAMESS (`DSDDH = FALSE`) and the DSD-DFTs branch is taken by setting `DSDDH = TRUE`, which calls `SUBROUTINE XCSCSFUN` and sets `DHFUNC` to `TRUE`. The latter keyword turns on the MP2 calculation on top of the GGA SCF.

If the desired exchange-correlation functional defined in `DFTYPE` corresponds to an optimized DSD-DFT, the coefficients of eq. 2.33 have tabulated values. However, in case the desired DSD-DFT is not optimized, GAMESS will need the user to define the various parameters required to run the DSD-DF in the `$DFT` group of the input file. In case of undefined parameters, GAMESS aborts with the following error message:

```
*** ERROR ***
THERE ARE NO DEFAULT VALUES FOR SCS-[FXFC]
PLEASE ENTER CHF, CGGA, CPARA, AND COPOS.
```

The coefficients `CHF`, `CGGA`, `CPARA`, and `COPOS` are double-precision variables. The keyword used to set up the empirical correction, `DCS6`, was already implemented into GAMESS.

The validation of the implementation was performed against Gaussian in which the DSD-DFT were initially implemented.

B. Computational details

Single point energy calculations were performed with our own version of GAMESS[14] 2012R1 on seven data sets with the def2-QZVPD[104] basis at various DFT[8;9] level: via nine generalized gradient approximation (GGA) DFs (BLYP,[66;67] BPBE,[66;71] BP86,[66;65] BPW91,[66;70] PBELYP,[71;67] PBE,[71] PBEP86,[71;65] PBEPW91,[71;70] OLYP[68;67]), one meta-GGA DF (tHCTH[141]), two hybrid DFs (B3LYP,[74] PBE0[76]), one meta-hybrid-GGA DF (tHCTHhyb[141]), two range-separated DFs (CAMB3LYP,[77] ωB97[142]), and ten double-hybrid DFs (B2PLYP,[143] DSD-BLYP,[13] DSD-BPBE,[13] DSD-BP86,[13] DSD-BPW91,[13] DSD-PBELYP,[13] DSD-PBEPBE,[13] DSD-PBEP86,[13] DSD-PBEPW91,[13] DSD-OLYP[13]), MP2,[4] spin-component-scaled MP2[10] (SCS-MP2), and CCSD(T).[60] The DSD-DF calculations were carried out with the same version of GAMESS, which includes these recent implementation.

The performance of the Def2-QZVPD basis set has already been addressed in an exhaustive basis set set study reported in section A. of the present Chapter.

At the DFT level, the army grade Lebedev[144] grid (NRAD=155, NLEB=1202) was used to solve the integrals. Both the D2[145] and the D3[146] empirical correction from S.Grimme were added to the standard DFT energy to correct for the missing long-range attraction.

At the MP2 level, second order perturbation energies were obtained on the unoccupied orbital space only (i.e. omitting the chemical core orbitals), and on the full orbital space (i.e. including the chemical core orbitals).

For the CCSD(T) calculations, the direct SCF[147;148] density convergence criteria was lowered to 2.5×10^{-7} (CONV=2.5D-07) and only integrals with small exponents were left out for the SCF (ICUT=11). The default values for the two electrons integrals transformation was lowered to CUTTRF=1D-11.

In either case, because the systems contain a large set of diffuse functions, all two-electron contributions of the Fock matrices are computed at each SCF step (FDIFF=FALSE).

The interaction energy E_{int} of all complexes was computed according to eq. 2.59.

$$E_{int} = E_C^{AB} - E_M^A - E_M^B \tag{2.59}$$

Where E_C^{AB} is the total energy of the complex AB. E_M^A and E_M^B are the total energies of the isolated monomers A and B, respectively.

C. Results and discussion

The overall performance of HF, reflected by it RMSD, is 2.598 kcal/mol. MP2 and SCS-MP2 are very close, with an RMSD of 0.673 and 0.665 kcal/mol, respectively. CCSD(T) gives an RMSD of 0.246 kasl/mol.

The root mean square deviations (RMSD) and the mean absolute errors (MAE) of the 59 methodologies presented herein are summarized in Tables 2.16, 2.17, 2.18, 2.19, 2.20 and displayed in Fig. 2.19 in the Appendix (section E.). For the sake of clarity, the overall performance (H panel of Fig. 2.19) is enlarged in Fig. 2.14.

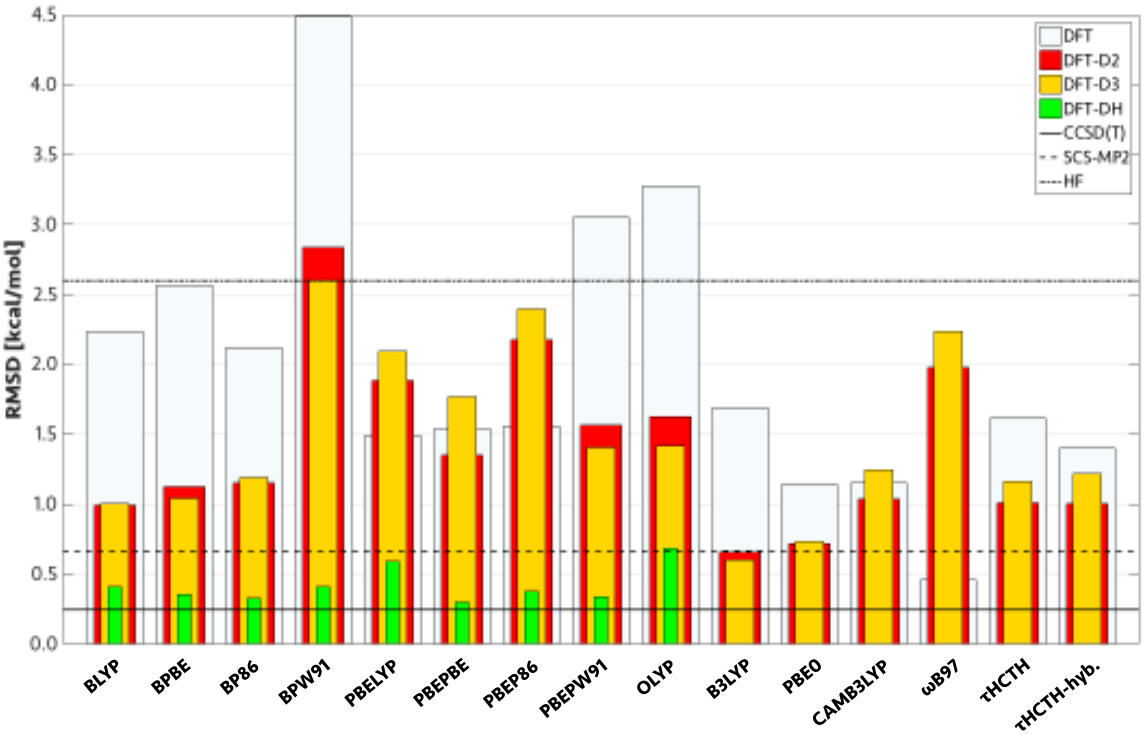


Figure 2.14: RMSD of the 56 methodologies. In white the standard DFTs (*i.e.* no dispersion correction), in red DFT-D2, in yellow DFT-D3 and in green the DSD-DFTs. The plain black line refers to CCSD(T), the dashed black line to SCS-MP2, and the dashed/dotted black line to HF. For the sake of clarity, MP2 result is not displayed (RMSD = 0.673 kcal/mol).

Overall performance of the methodologies

Fig. 2.19 points out the inability of four standard GGAs (*i.e.* without empirical corrections) to describe correlation effects. Indeed, BPBE, PBEPW91, PBEPW91 and OLYP have RMSD similar to

HF (dashed/dotted line), in the best case. Amongst the standard GGAs, PBELYP provides the best results with an RMSD=1.494 kcal/mol.

On average, non-GGA functionals show an accross-the-board improvement over standard GGAs with ωB97 describing the best non-covalent interactions: RMSD=0.4557 kcal/mol. It is to be mentioned that besides ωB97 the RMSD of the non-GGA functionals is rather similar to the best GGAs.

Not surprisingly, with a few exceptions, inclusion of empirical corrections improves the general performance of the various DFT (D2 correction in red and D3 correction in orange in Fig. 2.15). However, the performance of the best (non-)GGAs gets worse upon inclusion of empirical corrections with the largest deterioration in the case of ωB97: the D3 correction leads to an RMSD=2.236 kcal/mol. Similar trend is observed with PBELYP and PBEP86 (the best GGAs). In general, D3 tends to overestimate the non-bonding interactions although the variance gets smaller, as illustrated in Fig. 2.15 which plots reference (from Tables 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 2.7) against computed interaction energies. Panel A displays standard GGAs, panel B the GGA-D2, panel C the GGA-D3 and panel D the double-hybrid DSD-DFT. Climbing up the *stairway to heaven* from panel A to panel B generally improves the description of the weak part of the interaction energies. Panel B and C are similar even though a close look shows the trend of D3 correction to overbind: a slight left shift along the horizontal axis is noticed.

Inclusion of a perturbative treatment of the correlation energy results in a considerable improvement of the GGAs (green bars in Fig. 2.14, and panel D in Fig. 2.15). In term of RMSD, all double-hybrids sit between the SCS-MP2 (dashed line, Fig 2.14) and the state-of-the-art CCSD(T). DSD-PBEPBE even challenges the performance of coupled cluster (RMSD=0.299 kcal/mol and 0.246 kcal/mol, respectively). This drastic improvement is best shown in panel D of Fig. 2.15 where an (almost) perfect alignment of the 594 interaction energies with the reference line (depicted in black) is observed.

Performance on specific data sets

The necessity to design specific sets with a classification based the more dominant interaction amongst the non-covalent interaction makes sense. This is illustrated in Fig. 2.19 where the performance of the DFTs are evaluated on individual test sets. The accuracy of the investigated methods on specific stabilizing interactions is now investigated separately.

HB9. From panel A in Fig. 2.19, one can observe that half of the methodologies reported herein give higher accuracy on the hydrogen bonds than SCS-MP2 does (RMSD=0.802 kcal/mol). The worst results being obtained with BPW91, PBEPW91 and OLYP. Including the dispersion correction scheme, either via the D2 correction of the D3 correction, leads to more accurate prediction for only half of the methods, the other half being overestimated. The DSD-DFTs exceeds the SCS-MP2 accuracy with an RMSD as low as 0.233 kcal/mol for DSD-PBEPBE. Amongst the non-GGA functionals, PBE0 and CAM-B3LYP perform as well as the DSD-DFTs.

CT7/04. A disastrous situation for the standard DFTs is depicted in panel B of Fig. 2.19. Almost all the GGAs have an RMSD way above the accuracy of SCS-MP2 (RMSD=0.571 kcal/mol). Surprisingly inclusion of dispersion effects does not significantly affect the RMSD. However, a perturbative treatment of the correlation effects, as in the DSD-DFTs, has a considerable impact on the performance: half of the double-hybrids performs better than CCSD(T) (RMSD=0.349 kcal/mol) and the remainder sits between CCSD(T) and SCS-MP2.

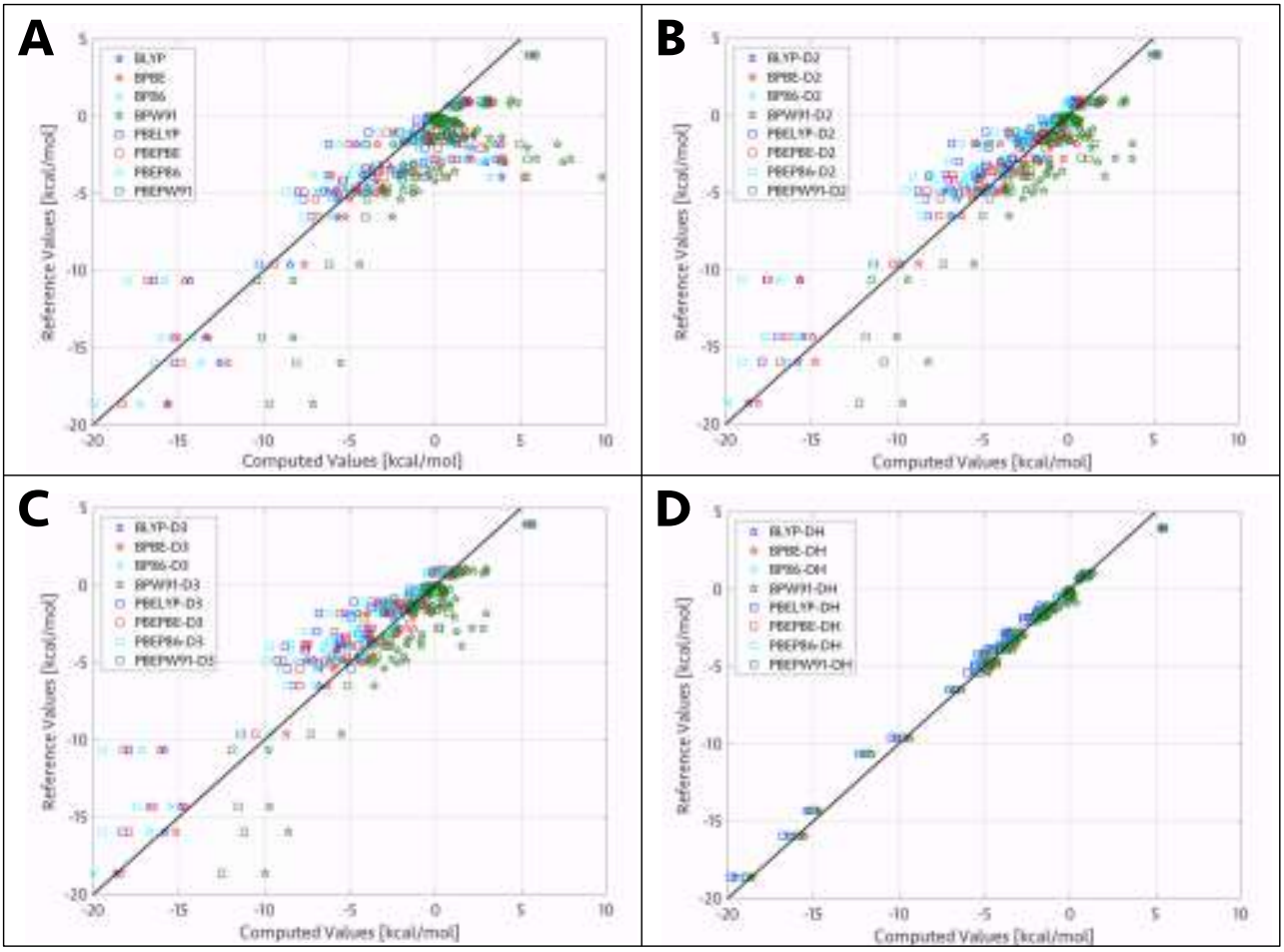
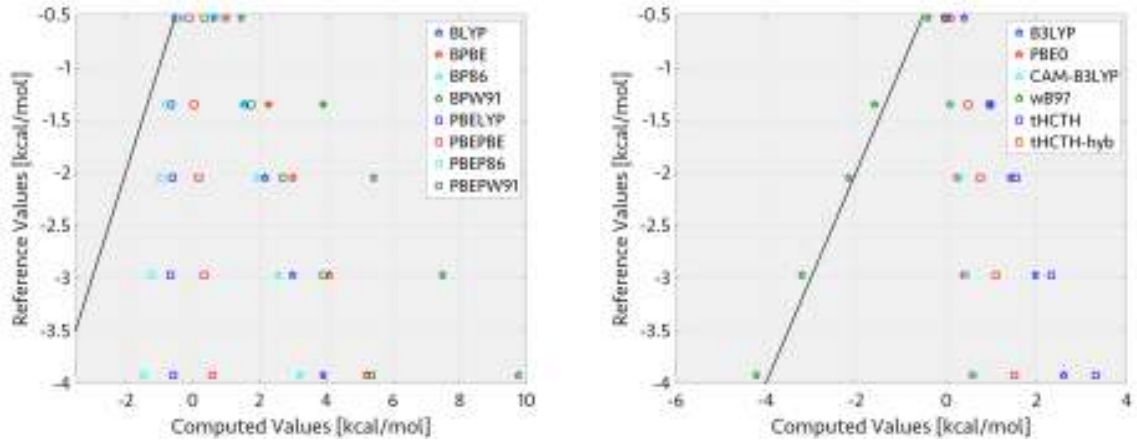


Figure 2.15: Climbing up the *stairway to heaven*. Correlation graph – reference values *vs.* computed values – of GGAs (A), GGAs-D2 (B), GGAs-D3 (C) and DSD-GGAs. The pentagrams refer to B88 exchange functional, the squares to PBE. In blue the LYP correlation functional, in red PBE, in cyan P86 and in green PW91.

DI9. The trend on the DI9 test set is very similar to the one observed on the HB9 set: empirical D2 and D3 corrections overestimate the interaction and the non-GGAs show an across-the-board improvement over the GGAs. In general, the DSD-DFTs provide a better accuracy than SCS-MP2 (RMSD=0.375kcal/mol), with DSD-PBEPW91 displaying the lowest RMSD (0.320 kcal/mol). For comparison, CCSD(T) has an RMSD of 0.192 kcal/mol.

ADIM5. Standard GGAs fail to describe the attracting interaction between alkane monomers. This observation is illustrated in Fig. 2.16. Note that subfigure (a) is an enlargement of panel A in Fig. 2.15. This observation holds for non-GGA functionals which also fail in the prediction of the interaction between alkanes (see Fig. 2.16 (b)). Only ωB97 somehow accurately describes the interaction energies and even outperforms MP2 and SCS-MP2. With either D2 or D3 correction, the missing dispersion term is recovered and the overall behavior of the functionals is improved, although overestimated. Panel F of Fig. 2.19 shows that the DSD-DFTs outperform SCS-MP2 to almost reach CCSD(T) accuracy.

IDISP4. The 56 methodologies give reliable predictions on the IDISP4 data set (panel E, Fig. 2.19). In the worst case, HF, BPW91 and OLYP have an RMSD between 1.373 and 1.576 kcal/mol. The best agreement is obtained with ωB97 (RMSD=0.703 kcal/mol), outperforming CCSD(T). The fact that standard functionals provide satisfying results is reflected when climbing along the *stairway to heaven*.



(a) GGA functionals. (b) Non-GGA functionals.

Figure 2.16: Correlation graph for the ADIM5 test set.

Indeed, inclusion of dispersion effects (via D2, D3 or DH) does not lead to noticeable improvements. Also, both post-HF method, *i.e.* MP2, SCS-MP2, and CCSD(T), gives RMSD=0.858, 0.829 and 0.744 kcal/mol, respectively. It is worth mentioning that MP2 and SCS-MP2 provides very similar results, suggesting that IDISP is governed by non-covalent interactions that takes place in the *strong* regime.

PPS11. Not surprisingly and as was the case for ADIM5, standard GGAs fail in describing the interaction energy of the PPS11 test set. Even though the RMSD is reasonable (RMSD between 1.445 kcal/mol and 5.726 kcal/mol) the interaction is opposite sign, as illustrated in Fig. 2.17. All 88 points predict repulsion between monomers. Note that Fig. 2.17 is an enlargement of panel A in Fig. 2.15. Non-GGA functionals compensate, somehow, for the wrong long range asymptotics, and indeed, most of the functionals predicts favorable interactions between monomers. Upon inclusion of empirical correction, through D2 or D3, the RMSD is considerably improved, but above all, the methodologies predict stabilizing interactions. Such shift is shown in panel B in Fig. 2.15. The DSD-DFTs predict very accurately the binding strength. Notably, besides DSD-BLYP, -PBELYP and -OLYP, the DSD-DFTs challenge the accuracy of CCSD(T). Interestingly, the RMSD summarized in table 2.20 points out the superior accuracy of SCS-MP2 over MP2.

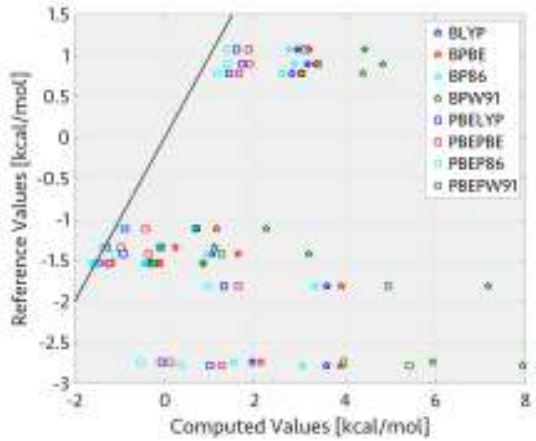
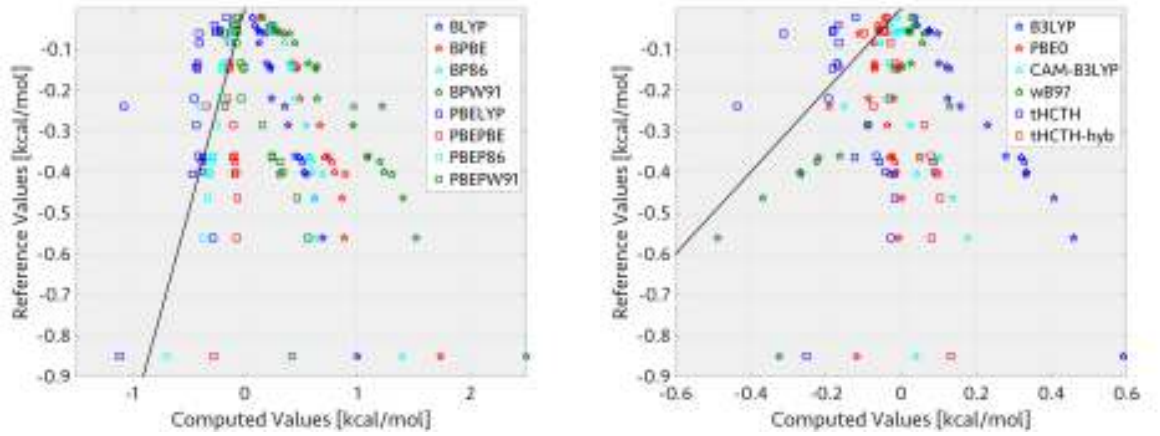


Figure 2.17: Correlation graph for the PPS11 test set. The pentagrams refer to B88 exchange functional, the squares to PBE. In blue the LYP correlation functional, in red PBE, in cyan P86 and in green PW91.

RG21. As was the case for ADIM5 and for PPS11, standard DFTs cannot describe the very weakly interacting rare gases. Fig. 2.18 reveals a dramatic situation for the GGAs (a) and the non-GGAs (b). The use of PBE exchange functional provides the right result for the wrong reason: it appears to recover for the missing dispersion interactions by over estimating the exchange energy. D2 and D3 corrections drastically improve the performance of the standard functionals. The DSD-DFTs yield very accurate results, in particular DSD-BLYP, -PBEPBE, and -PBEP86 which outperforms CCSD(T).



(a) GGA functionals.

(b) Non-GGA functionals.

Figure 2.18: Correlation graph for the RG21 test set.

D. Conclusion

The exhaustive performance study of methodologies on non-covalent interacting systems reveals a disastrous situation for the standard functionals. More importantly, some of the DFTs provide the correct results for the incorrect reason. Not surprisingly, the inclusion of dispersion effects via the empirical D2 or D3 scheme yields an overall improvement of correlation effects, although overestimated in most cases. The performance of the DSD-DFTs is brought to light: they often outperform SCS-MP2 and MP2 and, in a few cases, they challenge the accuracy of CCSD(T). Overall, the best performance is obtained with the DSD-PBEPBE double-hybrid functional, with an RMSD below 0.3 kcal/mol. DSD-BPBE, DSD-BP86 and DSD-PBEPW91 are within 0.05 kcal/mol difference on the RMSD. For comparison, CCSD(T) gives RMSD=0.246 kcal/mol. Both MP2 and SCS-MP2 are outperformed by the DSD-DFTs. Finally, it is worth noticing that the ωB97 (without empirical corrections) is the only functional competing with the DSD-scheme. As a matter of fact, with an RMSD=0.457 kcal/mol, it surpasses the accuracy of DSD-PBELYP and DSD-OLYP.

The double-hybrids, in the spin-component-scaled MP2 spirit appears as a promising technique to account for the incorrect long range asymptotics. However, the number of parameters to optimize clearly gives this method a semi-empirical character, although, it is quite different from the pure empirical approach frequently used to develop new functionals, *e.g.*, up to 58 parameters for the new MN15-L functionals.[149]

Nonetheless the $C_{c,o}$ and $C_{c,p}$ coefficients (eq. 2.33) should be independent of the functional choice. In this regard, new double-hybrids mixing the exact SCS-MP2 theory with the approximate GGA correlation are developed and reported in section VIII. Furthermore, aiming at minimizing the number of empirical parameters, the coefficients to be optimized are reduced from five fitted parameters to only two (C_{HF} and C_{GGA}, see eq. 2.63).

E. Appendix

Methods	HB9	CT7/04	DI9	ADIM5	IDISP4	PPS11	RG21	OVERALL
BLYP	1.741	1.891	1.442	4.994	1.049	3.276	0.747	2.236
	(1.466)	(1.446)	(1.383)	(4.423)	(0.887)	(2.770)	(0.601)	(1.584)
BPBE	2.150	1.709	1.932	5.906	1.018	3.539	1.022	2.564
	(1.973)	(1.199)	(1.805)	(5.284)	(0.855)	(3.077)	(0.854)	(1.879)
BP86	1.196	2.315	1.387	4.635	0.903	3.033	0.895	2.119
	(1.057)	(1.711)	(1.220)	(4.164)	(0.721)	(2.565)	(0.765)	(1.522)
BPW91	6.377	2.641	3.689	8.762	1.544	5.726	1.353	4.488
	(5.617)	(2.316)	(3.618)	(7.772)	(1.398)	(4.953)	(1.117)	(3.359)
PBELYP	0.646	3.477	0.469	1.959	0.909	1.726	0.273	1.494
	(0.598)	(3.213)	(0.292)	(1.574)	(0.706)	(1.127)	(0.212)	(0.879)
PBEPBE	0.512	3.094	0.709	2.775	0.871	1.922	0.243	1.544
	(0.387)	(2.652)	(0.607)	(2.381)	(0.660)	(1.429)	(0.177)	(0.932)
PBEP86	0.914	3.955	0.884	1.464	0.784	1.445	0.089	1.559
	(0.796)	(3.583)	(0.632)	(1.212)	(0.531)	(0.896)	(0.077)	(0.872)
PBEPW91	4.619	1.430	2.093	5.773	1.373	4.141	0.548	3.052
	(3.828)	(1.248)	(2.031)	(4.983)	(1.226)	(3.341)	(0.381)	(2.061)
OLYP	4.167	1.849	2.607	7.294	1.576	4.037	0.744	3.276
	(3.651)	(1.574)	(2.538)	(6.354)	(1.406)	(3.531)	(0.521)	(2.332)
B3LYP	0.907	0.766	1.129	4.135	1.044	2.503	0.585	1.692
	(0.719)	(0.525)	(1.090)	(3.642)	(0.845)	(2.144)	(0.461)	(1.134)
PBE0	0.393	1.163	0.624	2.795	0.938	1.580	0.293	1.143
	(0.309)	(0.756)	(0.569)	(2.414)	(0.689)	(1.269)	(0.205)	(0.701)
CAMB3LYP	0.554	0.454	0.622	2.848	0.974	1.715	0.373	1.150
	(0.466)	(0.311)	(0.578)	(2.480)	(0.726)	(1.449)	(0.271)	(0.735)
ωB97	0.844	0.438	0.474	0.208	0.703	0.311	0.190	0.457
	(0.780)	(0.335)	(0.372)	(0.197)	(0.414)	(0.244)	(0.157)	(0.323)
tHCTH	1.061	1.872	1.062	4.451	1.154	1.052	0.275	1.614
	(0.852)	(1.443)	(0.894)	(3.797)	(0.979)	(0.914)	(0.220)	(0.965)
tHCTH-hybrid	0.569	1.486	0.849	3.399	0.906	1.924	0.378	1.402
	(0.463)	(1.048)	(0.755)	(2.955)	(0.702)	(1.578)	(0.273)	(0.893)

Table 2.16: Root-mean-square deviation and mean absolute error (in parenthesis) of GGAs, hybrids, meta-GGAs and meta-hybrid-GGAs sitting on the first step of the *stairway to heaven*.

Methods	HB9	CT7/04	DI9	ADIM5	IDISP4	PPS11	RG21	OVERALL
BLYP-D2	0.428	2.710	0.463	0.763	0.581	0.685	0.176	0.993
	(0.294)	(2.279)	(0.304)	(0.650)	(0.520)	(0.496)	(0.123)	(0.526)
BPBE-D2	0.794	2.276	0.916	1.159	0.569	1.342	0.542	1.125
	(0.746)	(1.734)	(0.809)	(1.120)	(0.448)	(1.160)	(0.455)	(0.846)
BP86-D2	0.731	3.039	0.786	0.480	0.611	0.753	0.408	1.151
	(0.536)	(2.494)	(0.612)	(0.414)	(0.560)	(0.553)	(0.346)	(0.689)
BPW91-D2	4.854	2.102	2.458	3.988	0.779	3.526	0.863	2.838
	(4.255)	(1.856)	(2.410)	(3.609)	(0.620)	(3.036)	(0.718)	(2.151)
PBELYP-D2	1.775	4.236	1.538	2.849	0.633	0.875	0.637	1.885
	(1.711)	(3.976)	(1.434)	(2.589)	(0.572)	(0.792)	(0.569)	(1.394)
PBEPBE-D2	1.042	3.647	1.030	0.817	0.613	0.503	0.191	1.353
	(0.886)	(3.224)	(0.715)	(0.742)	(0.)486	(0.422)	(0.173)	(0.771)
PBEP86-D2	2.363	4.712	1.884	3.326	0.728	1.080	0.451	2.175
	(2.157)	(4.345)	(1.664)	(2.952)	(0.692)	(1.040)	(0.386)	(1.544)
PBEPW91-D2	3.117	1.513	0.880	0.997	0.670	1.988	0.154	1.563
	(2.467)	(1.291)	(0.823)	(0.820)	(0.447)	(1.424)	(0.135)	(0.955)
OLYP-D2	2.659	1.658	1.394	2.515	0.880	1.843	0.296	1.625
	(2.290)	(1.558)	(1.331)	(2.191)	(0.628)	(1.615)	(0.245)	(1.210)
B3LYP-D2	0.847	1.393	0.412	0.891	0.658	0.339	0.099	0.660
	(0.710)	(1.117)	(0.279)	(0.753)	(0.550)	(0.288)	(0.069)	(0.414)
PBE0-D2	1.045	1.552	0.626	0.099	0.707	0.380	0.081	0.717
	(0.864)	(1.183)	(0.408)	(0.084)	(0.470)	(0.299)	(0.063)	(0.404)
CAMB3LYP-D2	1.895	1.056	0.745	1.934	0.736	0.612	0.157	1.034
	(1.683)	(0.904)	(0.630)	(1.684)	(0.648)	(0.480)	(0.128)	(0.699)
ωB97-D2	2.360	1.041	1.616	4.967	0.955	2.083	0.360	1.980
	(2.141)	(0.982)	(1.518)	(4.312)	(0.889)	(1.893)	(0.249)	(1.379)
tHCTH-D2	0.767	2.557	0.738	0.371	0.632	0.354	0.319	1.008
	(0.534)	(2.087)	(0.530)	(0.366)	(0.418)	(0.314)	(0.295)	(0.591)
tHCTH-hybrid-D2	1.239	2.169	0.843	1.382	0.649	0.385	0.156	1.004
	(0.984)	(1.762)	(0.561)	(1.208)	(0.585)	(0.339)	(0.131)	(0.623)

Table 2.17: Root-mean-square deviation and mean absolute error (in parenthesis) of GGAs-D2, hybrids-D2, meta-GGAs-D2 and meta-hybrid-GGAs-D2 sitting on the second step of the *stairway to heaven*.

Methods	HB9	CT7/04	DI9	ADIM5	IDISP4	PPS11	RG21	OVERALL
BLYP-D3	0.233	2.853	0.631	0.612	0.736	0.356	0.116	1.006
	(0.172)	(2.477)	(0.416)	(0.510)	(0.436)	(0.308)	(0.098)	(0.491)
BPBE-D3	0.689	2.489	0.851	0.721	0.754	0.824	0.490	1.041
	(0.626)	(1.928)	(0.681)	(0.703)	(0.445)	(0.679)	(0.424)	(0.711)
BP86-D3	0.686	3.246	0.975	0.624	0.677	0.241	0.343	1.190
	(0.523)	(2.728)	(0.753)	(0.511)	(0.383)	(0.195)	(0.292)	(0.651)
BPW91-D3	4.740	1.826	2.105	3.554	1.136	3.006	0.810	2.602
	(4.198)	(1.572)	(2.048)	(3.192)	(0.988)	(2.555)	(0.687)	(1.965)
PBELYP-D3	1.833	4.494	1.909	3.285	0.726	1.324	0.677	2.094
	(1.768)	(4.247)	(1.796)	(3.006)	(0.427)	(1.273)	(0.600)	(1.593)
PBEPBE-D3	1.462	4.098	1.644	2.442	0.720	1.006	0.341	1.769
	(1.284)	(3.686)	(1.379)	(2.200)	(0.459)	(0.969)	(0.304)	(1.206)
PBEP86-D3	2.444	4.989	2.261	3.759	0.717	1.569	0.497	2.396
	(2.214)	(4.616)	(2.026)	(3.369)	(0.546)	(1.522)	(0.417)	(1.742)
PBEPW91-D3	2.990	1.501	0.567	0.594	0.998	1.520	0.125	1.407
	(2.410)	(1.255)	(0.461)	(0.466)	(0.815)	(0.970)	(0.105)	(0.804)
OLYP-D3	2.527	1.535	1.062	2.087	1.212	1.322	0.239	1.415
	(2.233)	(1.385)	(0.969)	(1.774)	(0.996)	(1.133)	(0.199)	(1.030)
B3LYP-D3	0.577	1.471	0.487	0.241	0.818	0.200	0.086	0.601
	(0.462)	(1.200)	(0.296)	(0.202)	(0.474)	(0.179)	(0.058)	(0.323)
PBE0-D3	0.961	1.632	0.721	0.221	0.814	0.274	0.084	0.733
	(0.788)	(1.281)	(0.468)	(0.219)	(0.454)	(0.238)	(0.068)	(0.413)
CAMB3LYP-D3	1.982	1.323	1.085	2.365	0.827	1.097	0.190	1.249
	(1.740)	(1.175)	(0.991)	(2.101)	(0.502)	(0.948)	(0.159)	(0.895)
ωB97-D3	2.461	1.340	1.995	5.397	0.833	2.615	0.407	2.236
	(2.199)	(1.253)	(1.880)	(4.729)	(0.743)	(2.374)	(0.276)	(1.576)
tHCTH-D3	0.799	2.791	1.054	0.801	0.865	0.679	0.351	1.164
	(0.592)	(2.347)	(0.805)	(0.783)	(0.569)	(0.658)	(0.326)	(0.755)
tHCTH-hybrid-D3	1.319	2.438	1.188	1.814	0.736	0.872	0.183	1.216
	(1.041)	(2.033)	(0.923)	(1.626)	(0.439)	(0.820)	(0.162)	(0.821)

Table 2.18: Root-mean-square deviation and mean absolute error (in parenthesis) of GGAs-D3, hybrids-D3, meta-GGAs-D3 and meta-hybrid-GGAs-D3 sitting on the third step of the *stairway to heaven*.

Methods	HB9	CT7/04	DI9	ADIM5	IDISP4	PPS11	RG21	OVERALL
DSD-BLYP	0.549	0.650	0.340	0.237	0.783	0.364	0.067	0.408
	(0.483)	(0.510)	(0.200)	(0.212)	(0.489)	(0.265)	(0.051)	(0.253)
DSD-BPBE	0.296	0.522	0.419	0.279	0.708	0.221	0.237	0.356
	(0.271)	(0.447)	(0.391)	(0.279)	(0.487)	(0.192)	(0.196)	(0.283)
DSD-BP86	0.222	0.513	0.351	0.169	0.735	0.155	0.221	0.326
	(0.156)	(0.311)	(0.273)	(0.165)	(0.479)	(0.129)	(0.190)	(0.215)
DSD-BPW91	0.272	0.503	0.405	0.261	0.710	0.215	0.455	0.411
	(0.242)	(0.417)	(0.375)	(0.260)	(0.487)	(0.187)	(0.366)	(0.325)
DSD-PBELYP	0.764	0.910	0.544	0.668	0.788	0.566	0.223	0.590
	(0.725)	(0.829)	(0.457)	(0.611)	(0.485)	(0.472)	(0.155)	(0.453)
DSD-PBEPBE	0.233	0.499	0.319	0.189	0.724	0.191	0.052	0.299
	(0.154)	(0.307)	(0.186)	(0.165)	(0.471)	(0.149)	(0.037)	(0.157)
DSD-PBEP86	0.402	0.698	0.394	0.227	0.755	0.221	0.034	0.377
	(0.358)	(0.537)	(0.228)	(0.200)	(0.459)	(0.184)	(0.022)	(0.218)
DSD-PBEPW91	0.266	0.501	0.320	0.152	0.730	0.200	0.270	0.338
	(0.190)	(0.302)	(0.184)	(0.136)	(0.468)	(0.156)	(0.192)	(0.209)
DSD-OLYP	0.756	0.854	0.615	1.317	0.786	0.708	0.184	0.686
	(0.685)	(0.687)	(0.503)	(1.164)	(0.584)	(0.593)	(0.137)	(0.501)
B2PLY	0.229	0.683	0.342	0.064	0.829	0.219	0.213	0.370
	(0.188)	(0.468)	(0.307)	(0.050)	(0.477)	(0.162)	(0.147)	(0.224)

Table 2.19: Root-mean-square deviation and mean absolute error (in parenthesis) of DSD-DF and B2PLY sitting on the fifth step of the *stairway to heaven*.

Methods	HB9	CT7/04	DI9	ADIM5	IDISP4	PPS11	RG21	OVERALL
HF	2.174	4.479	2.250	4.899	1.462	2.701	0.630	2.598
	(1.915)	(4.075)	(2.116)	(4.272)	(1.223)	(2.444)	(0.488)	(1.942)
MP2	0.113	0.565	0.439	0.242	0.858	1.430	0.048	0.673
	(0.093)	(0.437)	(0.267)	(0.203)	(0.499)	(0.978)	(0.042)	(0.317)
SCS-MP2	0.802	0.571	0.375	0.814	0.829	1.098	0.161	0.665
	(0.669)	(0.484)	(0.350)	(0.720)	(0.468)	(0.568)	(0.126)	(0.408)
CCSD(T)	0.091	0.349	0.195	0.123	0.744	0.135	0.106	0.246
	(0.056)	(0.296)	(0.126)	(0.098)	(0.398)	(0.094)	(0.071)	(0.126)

Table 2.20: Root-mean-square deviation and mean absolute error (in parenthesis) of HF, MP2, SCS-MP2 and CCSD(T).

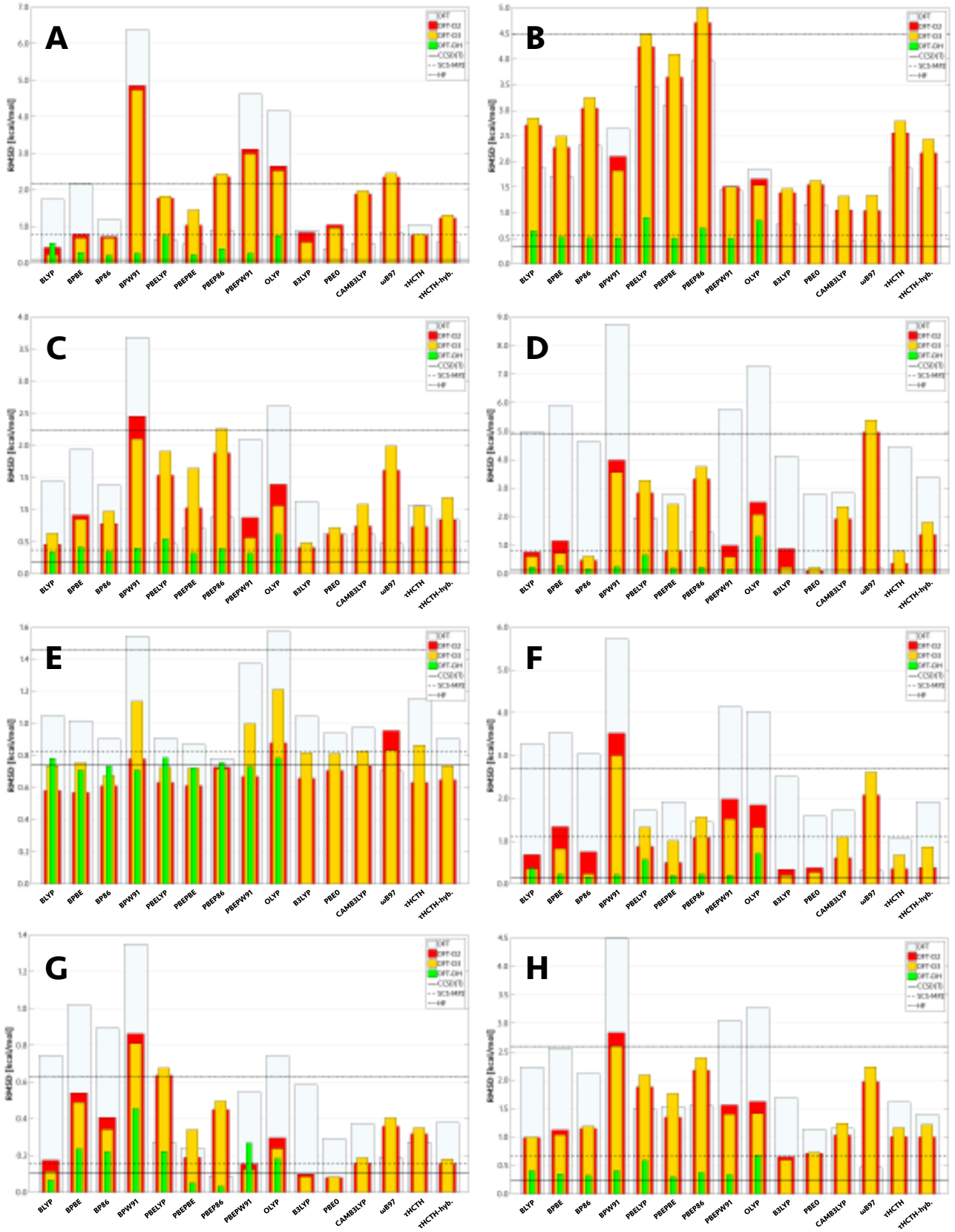


Figure 2.19: RMSD of the 56 methodologies on (A) HB9, (B) CT7/04, (C) DI9, (D) ADIM5, (E) IDISP4, (F) PPS11, (G) RG21, and (H) OVERALL. In white the standard DFTs (*i.e.* no dispersion correction), in red DFT-D2, in yellow DFT-D3 and in green the DSD-DFTs. The plain black line is CCSD(T), the dashed black line is SCS-MP2, and the dashed/dotted black line is HF. For the sake of clarity, MP2 is not displayed. Values from Tables 2.16, 2.17 , 2.18, 2.19, 2.20 were used to display panels A to H.

VIII Implementation, optimization and performance of spin-component-scaled DFTs

In the DSD-DFTs approach, the MP2 coefficients are functional- and basis set-dependent which makes this method of *empirical* character. One way to eliminate these drawbacks is to start from the derivation of the spin-component-scaled (SCS) MP2 from S. Grimme,[10] and to then extrapolate a (smaller) set of coefficients to the complete basis set limit (CBS). With the latter objectives as guideline, one makes sure that increasing the level of theory by slowly approaching the CBS, both (i) methods describing the exchange and, in particular, the correlation components of the total energy and (ii) the coefficients weighting the latter components are improved. The performance study carried out in the previous section showed the DSD-PBEPBE as a good candidate for the development of new double-hybrids families.

The SCS-MP2 coefficients were estimated on a few basic theoretical arguments. The derivation starts with the separation of antiparallel- and parallel-spin contributions to the total correlation energy as already suggested in the late 80s by G. A. Petersson *et al.*.[150;151].

$$E_c^{MP2} = E_{c,o}^{MP2} + E_{c,p}^{MP2} \tag{2.60}$$

where $E_{c,a}^{MP2}$ and $E_{c,p}^{MP2}$ are given by contribution pairs with antiparallel- (singulet), and parallel-spin (triplet), respectively. The systematic failures arising from weighting both contribution equally are corrected by a separate scaling, which leads to the expression of the SCS-MP2 correlation energy.

$$E_c^{SCS-MP2} = C_{c,o}E_{c,o}^{MP2} + C_{c,p}E_{c,p}^{MP2} \tag{2.61}$$

In his first argument, S. Grimme considered the fact that the correlation energy of two-electron systems with antiparrallel-spin contribution only (*e.g.* H_2 or He) is significantly underestimated by MP2: only 80% – 85% of the correlation energy is recovered. Consequently, $^6/_5$ appears as a reasonable opposite-spin scaling factor, $C_{c,o}$. Keeping in mind the fact that $E_c^{SCS-MP2}$ must equal E_c^{MP2}, the following relation must hold:

$$C_{c,p} = 1 - \frac{E_{c,o}}{E_{c,p}}\left(C_{c,o} - 1\right) \tag{2.62}$$

Because the ratio $^{E_{c,o}}/_{E_{c,p}}$ is considerably system-dependent, no precise estimate for $C_{c,p}$ can be obtained from theory. Consequently this parameter was determined empirically, at the QCISD(T)/QZV(3d2 f,2p1d) level of theory.

Merging the expression of SCS-MP2 to the approximate GGA exchange-correlation expression leads to a new family of double-hybrids, further referred to as SCS-DFTs. The formulation of their exchange-correlation expression reads:

$$E_{xc} = C_x E_x^{HF} + \left(1 - C_x\right) E_x^{GGA} + C_c E_c^{GGA} + \left(1 - C_c\right) \underbrace{\left[C_{c,o}E_{c,o}^{MP2} + C_{c,p}E_{c,p}^{MP2}\right]}_{SCS-MP2} \tag{2.63}$$

where $C_{c,o}$ and $C_{c,p}$ are the SCS-MP2 coefficient obtained from theory: $C_{c,o} = ^6/_5$, $C_{c,p} = ^1/_3$.

A. Implementation scheme into GAMESS

The previous implementation scheme of the DSD-DFTs (Fig. 2.13) was modified to include the new SCS-DFTs approach. The previous logical `DSDDH` keyword was changed to `XCDH`, referring to exchange-correlation double-hybrid. As was the case for the DSD-DFs, stand-alone GGA functional is the default

for `XCDH` (`XCDH = FALSE`). An additional keyword had to branch the DSD- and the new SCS-DFTs. From the user point of view, `DHCODE` ensures such branching.

The optimization of the coefficients is detailed in section B.

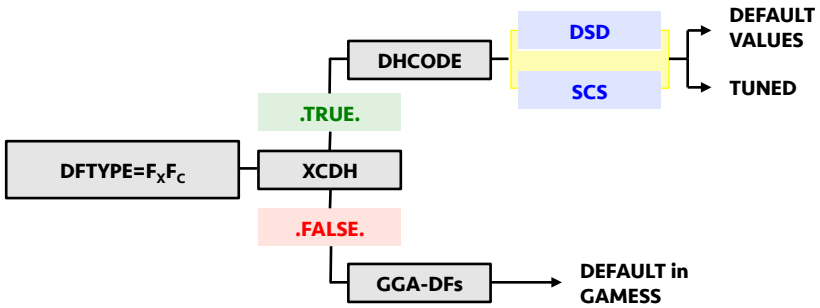


Figure 2.20: Implementation scheme of the SCS-DFs.

B. Optimization of the SCS-DFT coefficients

Single point energy calculations were performed with our own version of GAMESS[14] 2014R1 on the reduced data set displayed in Fig. 2.7 with various basis sets: the Dunning-style[112–114;126;152] cc-pVDZ, cc-pVTZ, cc-pVQZ and Def2-QZVP.[104] The army grade Lebedev[144] grid (NRAD=155, NLEB=1202) was used to solve the integrals.

The optimization of the parameters was carried out by a successive refinement of the 2-D grid, as depicted in Fig. 2.21, with C_x on the y-axis and C_c and the x-axis. In order to give similar contribution to each of the 13 systems composing the reduced set, a scaled root-mean-square deviation, $\mathrm{RMSD_{sc.}}$, was used in Fig. 2.21:

$$\mathrm{RMSD_{sc.}} = \sqrt{\frac{1}{13} \times \sum_{N_i=1}^{13} \left(\frac{E_{int,N_i}^{ref.} - E_{int,N_i}^{SCS-PBEPBE}}{E_{int,N_i}^{ref.}} \times 100 \right)^2} \tag{2.64}$$

In eq. 2.64 N_i is the N^{th} system of the reduced set (13 systems in total), $E_{int,N_i}^{ref.}$ is the reference interaction energy, as summarized in section IV, and $E_{int,N_i}^{SCS-PBEPBE}$ is the interaction energy obtained with the new SCS-PBEPBE double-hybrid.

Fig. 2.21 displays $RMSD_{sc.}$ on the final grid refinement. The blue region relates the smallest RMSD and the red region the largest RMSD. The best combination of C_{HF} and C_{GGA} is highlighted by a red cross. It is to be mentioned that a logarithmic scale was used to enhance the dynamic of $\mathrm{RMSD_{sc.}}$ in Fig. 2.21.

The extrapolation of the coefficients to the CBS limit follows a least-squares procedure fitting an exponential decay (see eq. 2.65), through the *trust-region* algorithm over the three points highlighted in Fig. 2.21. The extrapolation scheme from D. Feller was used:

$$f\left(L_{max}\right) = f_{CBS} + A \exp\left(-\alpha L_{max}\right) \tag{2.65}$$

Eq. 2.65 relates the highest angular momentum L_{max} to f_{CBS}, the value of the fitted parameter at the CBS limit. It is found that the RMSD is relatively insensitive to small variations in the parameters

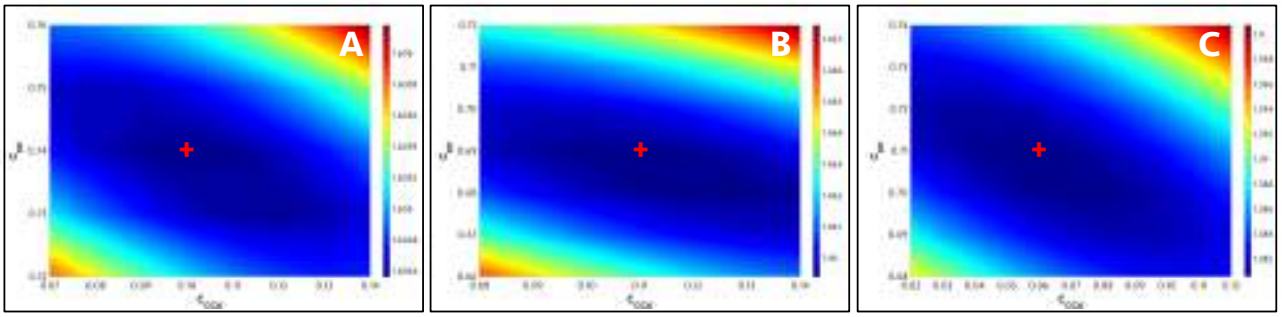


Figure 2.21: $RMSD_{sc.}$ on the reduced set at the cc-pVDZ (A), cc-pVTZ (B) and cc-pVQZ (C) level. Note that a logarithmic scale was used for an enhance dynamic. Highlighted by the red cross the best combination of C_{HF} and C_{GGA}.

around minimum, *e.g.*, ± 0.01 changes on the parameters yield a 0.041 kcal/mol difference on the RMSD. The final parameters adopted for the new SCS-PBEPBE are:

- $C_x = 0.71$ and $C_c = 0.06$

C. Performance evaluation of the new SCS-PBEPBE

The performance of the newly developed SCS-PBEPBE with $C_x = 0.71$ and $C_c = 0.06$ is assessed on the reduced set and on the PPS11 set at the def2-QZVPD level of theory and compared to the DSD-PBEPBE. Although SCS-PBEPBE can seem to be favoured over DSD-PBEPBE when applied to the reduced set, it is important to keep in mind the fact that the coefficients have not been optimized with the Ahlrichs basis sets and that the CBS coefficients are of higher quality than the quadruple-ζ coefficients. The RMSD and MAE are reported in Table 2.21.

methods	RMSD	MAE
SCS-PBEPBE	1.164	0.719
DSD-PBEPBE	0.934	0.505

(A) reduced data set

methods	RMSD	MAE
SCS-PBEPBE	0.369	0.254
DSD-PBEPBE	0.191	0.149

(B) PPS11

Table 2.21: (Unscaled) RMSD and MAE of SCS-PBEPBE and DSD-PBEPBE on the reduced set (A) and on the PPS11 set (B). Values are in kcal/mol.

With only two parameters, the new SCS-PBEPBE gives similar accuracy on the reduced set and on the PPS11 sets than the DSD-PBEPBE which counts five parameters. More importantly the total correlation energy arising from a combination of SCS-MP2 and the PBE correlation functional does not exceed 100%. This is ensured by using a single parameter weighting the correlation component of the double-hybrid functional. In the case of DSD-PBEPBE the sum of the C_c, $C_{c,o}$ and $C_{c,p}$ coefficients (eq. 2.33) is equal to 1.16, which over-counts for the correlation contribution to the total energy. Moreover, the empirical D2 correction added on top of the double-hybrid energy participates to further over-count correlation energy.

In comparison to DSD-DFTs in general, the new SCS-PBEPBE outperforms DSD-BLYP, DSD-PBELYP and DSD-OLYP (see table 2.19). Finally, it is somehow surprising that B2PLYP yields higher performance on the PPS11 set than the new SCS-PBEPBE (see table 2.19).

D. Conclusion and further development

With only two parameters, the SCS-PBEPBE functional performs as well as the five-parameter DSD-DFTs. Also the SCS-DF scheme does not over-weight the correlation energy component to reach high accuracy. Results on the remainder six sets (*i.e.* HB9, CT7/04, CI9, ADIM5, IDISP4 and RG21) would greatly contribute to assessing the performance of this new family of double-hybrids.

Further work would involve a basis set study to understand how the use of CBS coefficients could affect the results with smaller basis sets. It would also be of interest to use larger basis set (*e.g.* quintuple-ζ or sextuple-ζ) to evaluate the fit of the exponential decay reported herein. Last but not least, optimizing the parameters on larger test sets would provide more reliable coefficients on broader fields.

E. Appendix

First optimization round. Independently of the basis set used, the coefficients are set to $C_{HF} = \{0.1, 0.2, \ ... \ , 0.9\}$ and $C_{GGA} = \{0.1, 0.2, \ ... \ , 0.9\}$. This gives a general feeling of the potential energy surface at the SCS-PBEPBE level. The corresponding $\text{RMSD}_{\text{sc.}}$ are depicted in panels A, B, and C of Fig. 2.22.

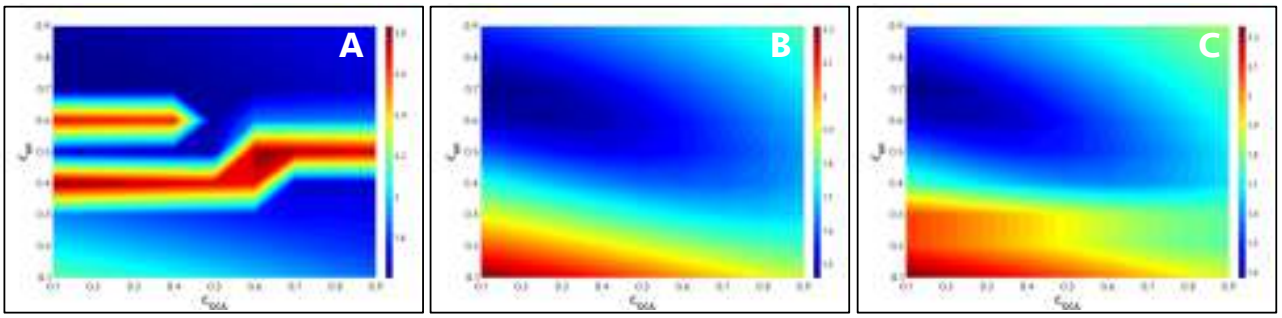


Figure 2.22: $\text{RMSD}_{\text{sc.}}$ on the reduced set at the cc-pVDZ (A), cc-pVTZ (B) and cc-pVQZ (C) level. Note that a logarithmic scale was used for an enhance dynamic.

Second optimization round. Based on the results obtained in the first round, the coefficients are refined with respect to the basis set used. The corresponding $\text{RMSD}_{\text{sc.}}$ are depicted in panels A, B, and C of Fig. 2.23.

- ▷ the coefficients with the cc-pVDZ basis set are refined to

$$C_{HF} = \{0.62, 0.64, 0.66, 0.68, 0.72, 0.74, 0.76, 0.78\}$$
$$C_{GGA} = \{0.12, 0.14, 0.16, 0.18, 0.22.0.24, 0.26, 0.28\}$$

- ▷ the coefficients with the cc-pVTZ basis set are refined to

$$C_{HF} = \{0.62, 0.64, 0.66, 0.68, 0.72, 0.74, 0.76, 0.78\}$$
$$C_{GGA} = \{0.02, 0.04, 0.06, 0.08, 0.12, 0.14, 0.16, 0.18\}$$

- ▷ the coefficients with the cc-pVQZ basis set are refined to

$$C_{HF} = \{0.62, 0.64, 0.66, 0.68, 0.72, 0.74, 0.76, 0.78\}$$
$$C_{GGA} = \{0.02, 0.04, 0.06, 0.08, 0.12, 0.14, 0.16, 0.18\}$$

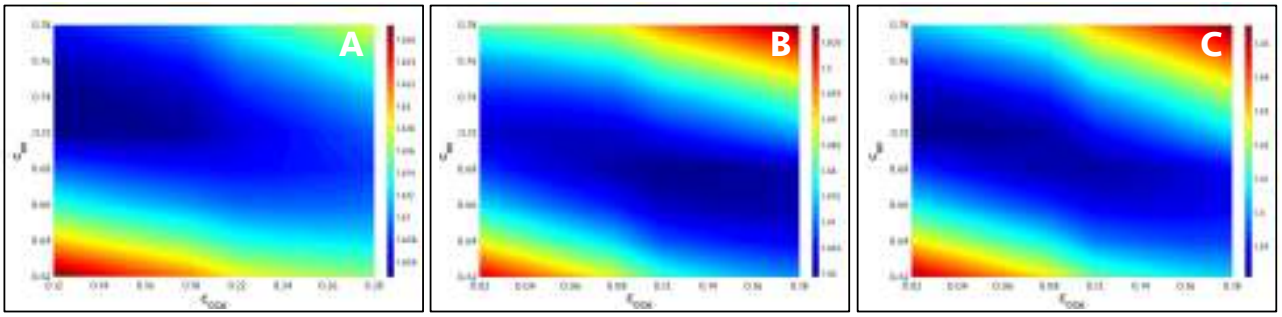


Figure 2.23: $RMSD_{sc.}$ on the reduced set at the cc-pVDZ (A), cc-pVTZ (B) and cc-pVQZ (C) level. Note that a logarithmic scale was used for an enhance dynamic.

Third optimization round. The coefficients are finally refined to following values. The corresponding $RMSD_{sc.}$ are depicted in panels A, B, and C of Fig. 2.21.

- ▷ the coefficients with the cc-pVDZ basis set are refined to

$$C_{HF} = \{0.72, 0.73, 0.74, 0.75, 0.76\}$$
$$C_{GGA} = \{0.07, 0.08, 0.09, 0.10, 0.11, 0.12, 0.13, 0.14\}$$

- ▷ the coefficients with the cc-pVTZ basis set are refined to

$$C_{HF} = \{0.66, 0.67, 0.68, 0.69, 0.70, 0.71, 0.72\}$$
$$C_{GGA} = \{0.08, 0.09, 0.10, 0.11, 0.12, 0.13, 0.14\}$$

- ▷ the coefficients with the cc-pVQZ basis set are refined to

$$C_{HF} = \{0.68, 0.69, 0.70, 0.71, 0.72, 0.73, 0.74\}$$
$$C_{GGA} = \{0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.10, 0.11, 0.12\}$$

IX Implementation and performance of the resolution-of-identity double-hybrid DFTs

The methodologies presented in sections VII (DSD-DFTs) and VIII (SCS-DFTs) have shown very good accuracy on non-covalent interactions, at high computational cost (*i.e.* cost of an MP2[4] run), warranting efforts to design methods that can reduce this cost. This section targets the development of cost-effective methods to account for correlation energy, without sacrificing accuracy.

In this aspect the resolution-of-identity[16;92–97] (RI), as introduced in sections III and VB. was merged to the DSD- and SCS-DFs.

A. Implementation scheme into GAMESS

The simplified implementation scheme in Fig. 2.20 was modified to allow GAMESS[14] users to run all double-hybrid schemes in the RI-approximation. This branching was made possible by introducing a new logical keyword: `RIDH`. The default was set to `FALSE` which computes the four-index two-electron integrals in its standard formulation (see eq. 2.15). When `RIDH` is set to `TRUE`, the RI-V method from Whitten[16] is used to reduce the cost: the $\langle ij|kl\rangle$ are approximated by the three- and two-index two-electron integrals.

$$\langle ij|kl\rangle \approx \sum_{\mu}\sum_{\nu} \langle ij|\mu\rangle \langle \mu|\nu\rangle^{-1} \langle \nu|kl\rangle \tag{2.66}$$

As already mentioned in section VB., different subroutines of GAMESS were modified and in particular the H & I shell common blocks and loops to gather information on the H & I shells were added, and the argument size storage was tested and validated.

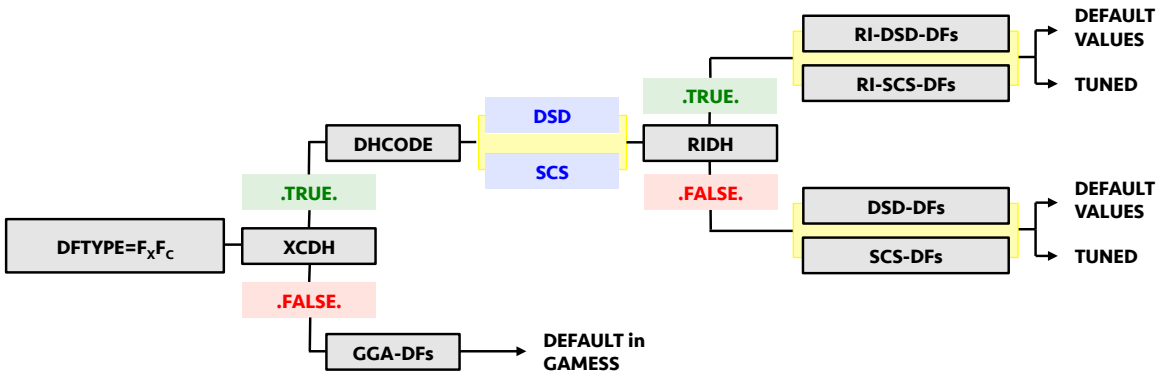


Figure 2.24: Implementation scheme of the RI-DH.

Hereafter, the results of the extensively studied RI-DSD-DFTs are presented.

B. Computational details

Single point energy calculations were performed with our own version of GAMESS[14] 2014R1 on the seven data sets displayed in Fig. 2.6 with the Def2-QZVPD[104] as main basis set and Def2-QZVP[153] as auxiliary basis set. The quality of the setup was exhaustively studied in section VB. The army grade Lebedev[144] grid (NRAD=155, NLEB=1202) was used to solve the integrals.

The performance of RI-DSD-BLYP, RI-DSD-BPBE, RI-DSD-BP86, RI-DSD-BPW91, RI-DSD-PBELYP, RI-DSD-PBEPBE, RI-DSD-PBEP86, and RI-DSD-BPW91 was assessed against the corresponding stand-alone DSD-DF schemes. Notably, RMSDs and MAEs were computed as:

$$RMSD = \sqrt{\frac{1}{N} \times \sum_{i=1}^{N} \left(E_{int,i}^{DSD-DF} - E_{int,i}^{RI-DSD-DF} \right)^2} \tag{2.67}$$

$$MAE = \frac{1}{N} \sum_{i=1}^{N} \sqrt{\left(E_{int,i}^{DSD-DF} - E_{int,i}^{RI-DSD-DF} \right)^2} \tag{2.68}$$

where $E_{int,i}^{DSD-DF}$ and $E_{int,i}^{RI-DSD-DF}$ are the interaction energy of system i of the stand-alone DSD-DFT and of the new cost-effective RI-DSD-DFT, respectively.

C. Results and discussion

The correlation graph in Fig. 2.25 displays the interaction energies of the nine double-hybrids investigated on the seven data sets (*i.e.* HB9, CT7/04, DI9, ADIM5, IDISP4, PPS11, and RG21).

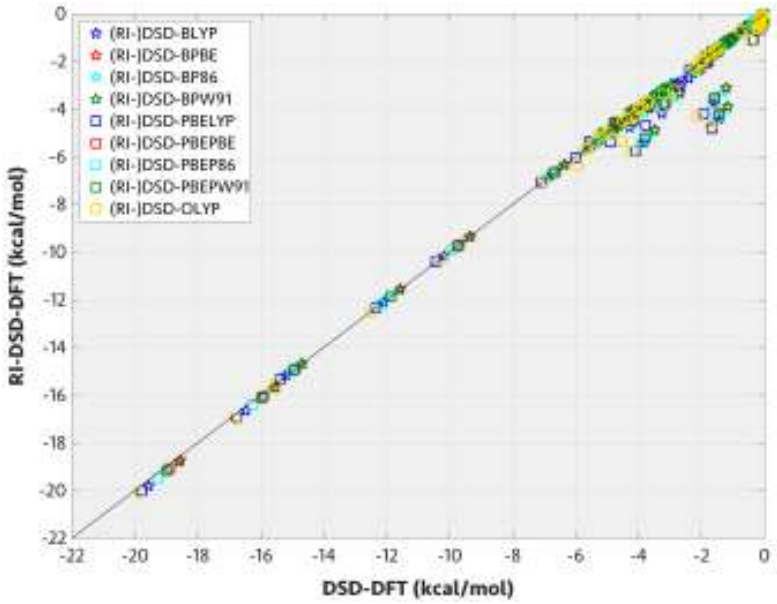


Figure 2.25: Correlation graph: interaction energies of DSD-DFTs against RI-DSD-DFT on the seven data sets.

Fig. 2.25 presents the performance of the RI-DSD-DFT scheme. Among the 594 computed interaction energies, only the difficult case of PPS11 shows RMSDs higher than 0.6 kcal/mol. Nonethless, a strong correlation of $R^2 = 0.97127$ for the relationship between the DSD-DFs and the RI-DSD-DFs is obtained, along with an RMSD between 0.471 and 0.577 kcal/mol over the seven validation sets. Overall the results indicate that the interaction energies obtained at both levels of theory are in excellent agreement. Such agreement is comparable to standard RI-MP2 results based on the HF orbitals.[154;155] However, the largest RMSD and MAE (summarized in table 2.23 in the Appendix) is 1.778 and 0.537 kcal/mol, respectively, with DSD-PBELYP on the PPS11 test set. Considering that the interacting energies on that particular set span +1.074 to -2.780 kcal/mol, this reflets some limitations of the RI-approximation. Although in term of absolute number, 0.537 kcal/mol of MAE is reasonable, such an error on an interaction energy of 1.074 kcal/mol represents a non-negligible 50% error.

It has to be mentioned that the chemical systems investigated are critical systems in the sense that dispersion effects are very strong. Therefore, the error introduced by approximating the four-index two-electron transformations is expected to be larger than on the other test sets.

Finally, taking into consideration that the RI-DSD-DFs are significantly faster than the DSD-DFs (51.5 times faster on average), RI-DSD-DFs are concluded to be well-suited as accurate cost-effective methods for the treatment of systems governed by correlation effects. Table 2.22 illustrates the computational gain of using the RI-DSD-DF over the DSD-DFs.

Methods	HB9	CT7/04	DI9	ADIM5	IDISP4	PPS11	RG21	OVERALL
DSD-DFs	6-22:39	1-13:18	5-3:27	32-9:9	10-6:48	98-1:37	1-1:31	155-10:33
RI-DSD-DFs	0-2:51	0-0:47	0-2:13	1-2:51	0-5:49	1-9:21	0-0:13	3-0:25
RATIO	58.3	47.1	55.5	28.9	42.4	70.6	48.6	51.5

Table 2.22: Overall CPU timing for the nine DSD-DFs and for the nine RI-DSD-DFs. Values are in written in a Days-Hours:Minutes format.

An overall speed up of 51.5 is obtained. in term of computational time, the total CPU time required to compute the 594 interaction energies is 155 days at the DSD-DFTs level. Upon the RI-approximation the CPU time drops to (only) 3 days.

D. Conclusion

The RI-DSD-DF methods give an accurate description of the correlation effect and results are in good agreement with the standard DSD-DFs. Although the interaction energies obtained with the RI-MP2 methodology slightly differ from those evaluated with the exact MP2, an important saving in computational time and system requirements such as, *e.g.*, disk space and memory is achieved.

E. Appendix

Methods	HB9	CT7/04	DI9	ADIM5	IDISP4	PPS11	RG21	OVERALL
DSD-BLYP	0.103	0.079	0.532	0.498	0.217	1.120	0.001	0.533
	(0.057)	(0.044)	(0.203)	(0.377)	(0.181)	(0.516)	(0.001)	(0.174)
DSD-BPBE	0.076	0.064	0.468	0.325	0.167	1.012	0.001	0.471
	(0.038)	(0.035)	(0.173)	(0.232)	(0.131)	(0.471)	(0.001)	(0.143)
DSD-BP86	0.084	0.069	0.491	0.376	0.166	1.036	0.002	0.486
	(0.043)	(0.038)	(0.185)	(0.278)	(0.141)	(0.471)	(0.001)	(0.150)
DSD-BPW91	0.077	0.064	0.466	0.332	0.165	1.007	0.203	0.481
	(0.039)	(0.035)	(0.173)	(0.238)	(0.127)	(0.467)	(0.145)	(0.184)
DSD-PBELYP	0.106	0.083	0.557	0.488	0.205	1.178	0.280	0.577
	(0.060)	(0.046)	(0.213)	(0.378)	(0.176)	(0.537)	(0.205)	(0.238)
DSD-PBEPBE	0.078	0.067	0.484	0.319	0.171	1.048	0.001	0.486
	(0.041)	(0.036)	(0.180)	(0.235)	(0.143)	(0.481)	(0.001)	(0.148)
DSD-PBEP86	0.086	0.072	0.507	0.370	0.165	1.073	0.001	0.502
	(0.046)	(0.040)	(0.191)	(0.279)	(0.137)	(0.485)	(0.001)	(0.154)
DSD-PBEPW91	0.078	0.066	0.474	0.320	0.164	1.025	0.208	0.489
	(0.041)	(0.036)	(0.177)	(0.237)	(0.136)	(0.470)	(0.152)	(0.188)
DSD-OLYP	0.099	0.078	0.520	0.405	0.197	1.119	0.264	0.543
	(0.056)	(0.043)	(0.198)	(0.307)	(0.172)	(0.521)	(0.192)	(0.222)
AVERAGE	0.087	0.071	0.500	0.382	0.180	1.069	0.107	0.508
	(0.047)	(0.039)	(0.188)	(0.284)	(0.149)	(0.491)	(0.078)	(0.178)

Table 2.23: Root-mean-square deviation and mean absolute error (in parenthesis) of RI-DSD-DFs.

X Implementation of $erfc$ Møller-Plesset second order perturbation theory

On the way towards the implementation of a different approach to describe the correlation component of the total energy in the double-hybrid DFT, the attenuated MP2 developed by M. Head-Gordon *et al.*[15;156] was implemented in GAMESS, so that users can select between three types of MP2 types: (i) standard MP2, (ii) spin-component-scaled (SCS) MP2 and (iii) range-separated MP2, further referred to as MP2($erfc$), which is based on the error function and its complement.

The remainder of this section starts with a background on MP2($erfc$). Then, the implementation of MP2($erfc$) in GAMESS is detailed. The theory, including range-separation and integrals within the $erfc$ frame, was inspired from Refs. [15; 156–164].

A. Background

MP2 is one of the simplest ways to account for correlation energy. During the last few years, MP2 has been at the origin of new WFT development. Even though MP2 performs excellently for some types of interactions, such as hydrogen bonds, it provides an inadequate description of the weak intermolecular interactions.[33] Notably, MP2 overestimates interaction energies. Interaction of benzene dimers,[34;165] DNA base pairs and amino acids pairs,[35] are examples of the poor MP2 behavior. As already mentioned, S. Grimme suggested a cost-free modification known as SCS-MP2,[10] which considerably improved the description of correlation energy. More recently, M. Head-Gordon *et al.* came up with an attenuated Coulomb operator, leading to a range-separated MP2.[15] In the latter method, the traditional Coulomb operator was partitioned into two parts:

$$\frac{1}{R} = \frac{erfc\left(\omega \times R\right)}{R} + \frac{erf\left(\omega \times R\right)}{R} \tag{2.69}$$

The first term of eq. 2.69 is singular and short-range, while the second term is non-singular but long-range. As depicted in Fig. 2.26, this separator function yields both a very rapidly decaying short-range component and a smooth long-range component.

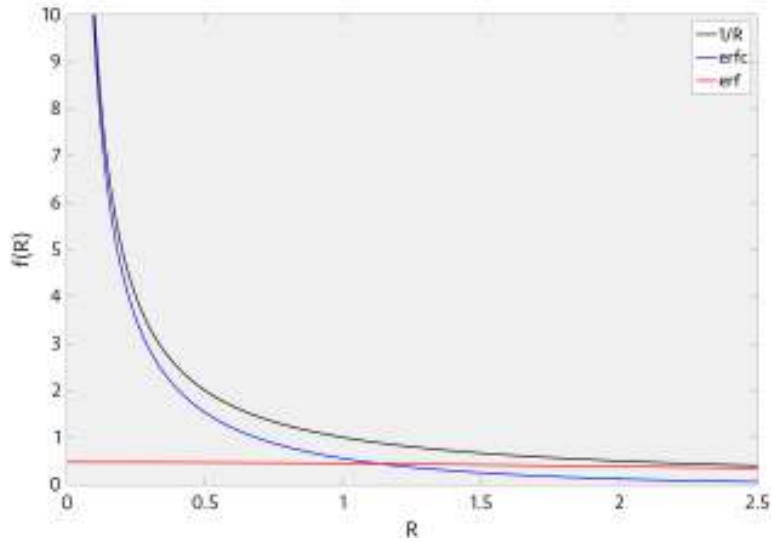


Figure 2.26: In black the Coulomb operator $^{1}/_{R}$, in blue $^{erfc(\omega \times R)}/_{R}$, and in red $^{erf(\omega \times R)}/_{R}$. The attenuation parameter ω is set to 0.420 Å^{-1} as reported by M. Head-Gordon *et al.*[15]

Since long-range contributions control the overall computational cost of traditional MP2 calculations and also limit their accuracy, the short-range MP2($erfc$) considers only the Coulomb-attenuated part by excluding the background ${}^{erf(\omega \times R)}/_{R}$. In this context, the integral I in eq. 2.35 reads

$$\begin{aligned} I &= \int\int d\mathbf{r_1} d\mathbf{r_2} \chi_a(\mathbf{r_1}) \chi_b(\mathbf{r_1}) \frac{erfc(\omega \times R)}{R} \chi_c(\mathbf{r_2}) \chi_d(\mathbf{r_2}) \\ &= \left\langle \chi_a \chi_b \left| \frac{erfc(\omega \times R)}{R} \right| \chi_c \chi_d \right\rangle \end{aligned} \tag{2.70}$$

which can be reduced to a one-dimensional integral in inverse space:

$$I = \frac{1}{2(\pi^2 R)^2} \int_0^\infty du \ \sin(u) \left(\int_0^\infty dr \ \sin\left(\frac{u}{R} r\right) F(r) \right) exp\left[\frac{\frac{u^2}{p} + \frac{u^2}{q}}{4R^2} \right] \tag{2.71}$$

with

$$F(r) = \frac{erfc(\omega \times R)}{R} \tag{2.72}$$

The integral I in eq. 2.71 depends on the inter-particle distance R, on the exponents p and q of the original basis function quartet, and most importantly, on R, the distance between the two distributions. The attenuation parameter ω was optimized at the cc-pVDZ level on existing data sets (S22,[99] S66,[166] and P76[167]) and was found to equal 0.420 $Å^{-1}$.[15]

B. Implementation into GAMESS

MP2($erfc$) is implemented in a two-step process in the DDI version of MP2. First, the MP2($erfc$) energy was obtained by taking out the the background from the standard MP2 energy using the relation $erf + erfc = 1$. This requires two MP2 runs:

$$\text{MP2}(erfc) = \text{MP2} - \text{MP2}(erf) \tag{2.73}$$

Where MP2(erf) is evaluated with

$$\frac{1}{R} \approx \frac{erf(\omega \times R)}{R} \tag{2.74}$$

The error function erf was already implemented in the DFT-code of GAMESS and was used as starting point for computing the attenuated four-index two-electron integrals. Second, the differentiation expressed in eq. 2.73 is performed *on-the-fly*, in a single MP2 run. For the sake of clarity, the two-step process is detailed below, and the one-step process is briefly explained afterwards.

To minimize the number of modifications, all atomic orbital integrals were evaluated with the `HONDO-RYS` package (selecting `INTTYP=RYSQUAD`, in `$CONTROLL`, see section II for more information). The `LRINT` variable was used to branch standard integrals and erf-style integrals. In the case of a single-point MP2 energy evaluation (see Fig. 2.5), the first run had `LRINT` set to `FALSE` (to collect the MP2 energy) in `SUBROUTINE WFNMP2` (gamess.src), and then, in a second run `LRINT=TRUE` (to collect the MP2(erf) energy).

```
IF(CODEMP.EQ.DDI) THEN
   CALL MP2DDI
   LRINT=.TRUE.
   CALL MP2DDI
   LRINT=.FALSE.
ENDIF
```

The flow-diagram of a single-point MP2 calculation (Fig. 2.5) shows that the values of LRINT and all associated variables to the error function have to be propagated in the subroutines MP2DDI, PARTRAN, S0000, and GENRAL. This was done by adding the following common block and variables to the above-mentioned subroutines.

```
COMMON /NLRC / LCFLAG,RS,LRINT,EMU,EMU2,LRFILE
LOGICAL LRINT,RS,LCFLAG
```

Using the HOND-RYS package all of the integrals were done in the source file int2a.src. In the first run all the coefficients of the recurrence relations required to solve the Rys quadrature are unattenuated. In the second run, an attenuation of both ρ, a quantity associated with a two-dimensional integral, and $F00$, the fundamental integrals is required:

```
IF(LRINT) THEN
   RHO0 = RHO
   RHO = RHO0 × EMU2/(RHO0+EMU2)
   F00 = F00 × SQRT(EMU2/(RHO0+EMU2))
ENDIF
```

where RHO refers to ρ and EMU2 is the squared attenuation parameter ω^2.

Once the two runs, *i.e.* MP2 and MP2(erf), are carried out, the final energies of each run are collected and combined to print only the MP2($erfc$) energy. For the *on-the-fly* calculation, a new keyword logical variable was defined and added to the common block. This new variable RS ensures branching between standard MP2 and single run MP2($erfc$). The latter IF condition was modified to allow *on-the-fly* evaluation of the $erfc$-style integrals:

```
IF(LRINT.OR.RS) THEN
   RHO0 = RHO
   EFR = EMU2/(RHO0+EMU2)
   IF(LRINT) THEN
      RHO = RHO0 × EFR
      F00 = F00 × SQRT(EFR)
   ELSE THEN
      RHO = RHO0 × ONE-EFR
      F00 = F00 × ONE-SQRT(EFR)
   ENDIF
ENDIF
```

Last but not least, the implementation of the $erfc$-style integrals were extended to the Pople-Hehre algorithm (see section II for more information). All in all, GAMESS users can choose between three types of MP2 types by selecting a combination of (new) keywords, as illustrated in Fig. 2.27.

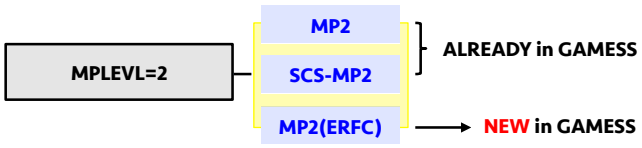


Figure 2.27: New implementation scheme of MP2.

XI Implementation of the double-hybrid MP2($erfc$) DFTs

In the previous approaches, *i.e.*, DSD- and SCS-DFs, the correlation component of the double-hybrid was based on scaling differently the same- and opposite-spin contributions to the correlation energy to improve the performance of MP2. In this section, another double-hybrid family, based on the attenuated MP2 as described in the previous section (section X), is presented. The general expression of the exchange-correlation functional reads:

$$E_{xc} = C_x E_x^{HF} + (1 - C_x) E_x^{GGA} + C_c E_c^{GGA} + (1 - C_c) E_c^{MP2-erfc} \tag{2.75}$$

where $E_c^{MP2-erfc}$ is the correlation energy obtained with the attenuated MP2 pertubational (section X) term based on the Kohn-Sham orbitals. This new family is referred to as DH($erfc$)-DFs.

A. Implementation scheme into GAMESS

In a final step, the implementation scheme shown in Fig. 2.24 was modified to lead to the one depicted in Fig. 2.28. In this final implementation scheme, `ERFC` is added to the list of double-hybrid schemes.

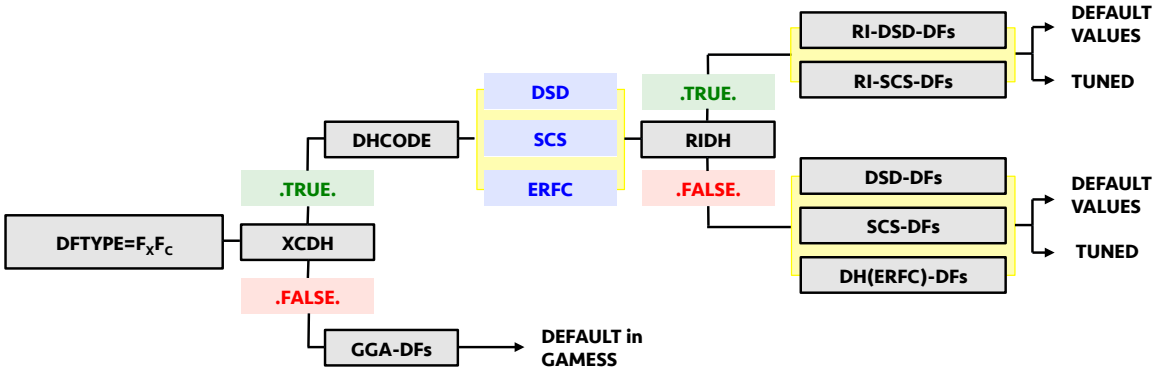


Figure 2.28: Implementation scheme of the DH-DF($erfc$).

B. Conclusion and further work

The successful implementation of a new family of double-hybrid is reported here. It mixes attenuated MP2 and approximate correlation energy within a standard hybrid frame.

Optimization of the new DH($erfc$)-DFs would start at the cc-pVDZ and at the cc-pVTZ basis set level, since optimized ω values are reported in the literature.[15;156] Further work would involve the optimization of the new DH($erfc$)-DFs, following the procedure described in the section VIII. This effort would involve the CBS extrapolation of (i) the ω parameter and (ii) the fitted parameters C_x and C_c. The RI version of DH($erfc$)-DFs would lead to working out the attenuated three- and two-center integrals

$$\langle ij|kl\rangle \approx \sum_{\mu}\sum_{\nu} \langle ij|\mu\rangle \langle \mu|\nu\rangle^{-1} \langle \nu|kl\rangle \tag{2.76}$$

with the Coulomb attenuated operator:

$$\frac{1}{r} \approx \frac{erfc(\omega \times r)}{r} \tag{2.77}$$

Chapter 3

Theory and experiment: a synergy towards understanding and predicting chemistry

I Introduction

Among weak noncovalent interactions, van der Waals (vdW) interactions with aromatic moieties have taken a prominent role due to their importance in both biological and chemical processes.[168;169] With large aromatic hydrocarbons (PAH) or carbon nanostructures involved, these vdW forces are mostly accounted for by π-π interactions.[170;171] Triggered by the discovery of fullerene C_{60},[172] studies on non-planar PAHs added the entity of curvature to the concept of aromaticity.[173;174] A large fraction of these PAHs displayed concave-convex π-π interactions in their respective crystal structures.[175–179] Consequently, their shape complementarity to the convex fullerene has been utilized to study concave-convex π-π interactions with the latter in solid state.[180–184] The assembly with corannulene, the smallest curved fragment of C_{60},[173;174;185] was used early-on as a model system. The interaction was believed to be only present in gas phase,[186] until evidence was found for its existence on surface[187] and solid state.[182] Up-to-date, there have been only a handful of examples of corannulene-based complexations with fullerene in solution.[188–194] Involved were either heterosubstituted corannulenes[188–190] or molecular tweezers consisting of two corannulene subunits.[191–194]

Pentaindenocorannulene[195] (PIC), a $C_{50}H_{20}$ curved polycyclic aromatic hydrocarbon (PAH), is shown to form concave-convex π-π interactions with C_{60}. Experimental evidence for their existence in solid state and in solution is given, supporting the high-level computations presented hereafter. The PAH itself is shown to be highly prone to self-assemble. Furthermore, a new polymorph motif following a columnar-like stacking is presented.

Note that this chapter is an adapted version of the recently published communication *Pentaindenocorannulene: Properties, assemblies & C60 complex.*[196]

II Computational details

The structural and energetic analysis of the molecular systems described in this study were carried out using the GAMESS 2014R1 software.[14] All geometries were optimized in the gas phase with DFT methods.[8;9] Comparisons were made across a variety of density functional types, including B97,[197] PBE,[71;123] PBE0,[76] TPSS.[198;199] Final structural comparisons were made at the B97-D[106]/6-311G(2d,p)[200;201] level of theory. Full geometry optimizations were performed and uniquely characterized via second derivatives (Hessian) analysis to determine the number of imaginary frequencies (0=minima; 1=transition state) as well as zero point and thermal corrections. The gradient convergence tolerance and the root mean square gradient were lowered to 0.00001 Hartree/Bohr and to 0.000003 Hartree/Bohr, respectively. In all calculations, a larger number of radial points in the Euler-MacLaurin quadrature and a finer Lebedev[144] grid than the army-grade grid were used (nrad=155 and nleb=1202, respectively). The SCF density convergence criterion was lowered to 2.5×10^{-07}, and the integral cutoff was lowered to 10^{-11}.

Further single point calculations were carried out at the B97-D/Def2-TZVP[104]//B97-D/6-311G(2d,p) level of theory for more accurate energetic and property information. The basis set choices were validated by comparing the interaction energy of representative structures as obtained at the B97-D/Def2-TZVP with the B97-D/CBS limit.[124]

Effects of solvation were account for using the COSab[202;203] modified conductor like screen model, with the dielectric constant for ortho-dichlorobenzene (structural analysis), THF (electrochemical data)

1,1,2,2-tetrachloroethane (NMR) and chloroform (UV-Vis data). The multiplicative factor for van der Waals radii used for cavity construction was 1.3 to account for proper solute-solvent interaction, and the outlying charge error was treated with the double-cavity method. UV-Vis data was determined in chloroform using TD-DFT[43;204–210]/CAMB3LYP[77]/Def2-TZVP//B97-D/6-311G(2d,p), with the Gaussian package.[211] Further details are provided in the the corresponding section. NMR data was determined in 1,1,2,2-tetrachloroethene (TCE) using CSGT[212;213]/B97-D/Def2-TZVP//B97-D/6-311G(2d,p), with the Gaussian package.

Reduction potentials were determined at the B97-3[75]/Def2-TZVPD//B97-D/6-311G(2d,p) level using $E^{\circ} = {}^{-\Delta E}/_{nF}$, where n=1, 2, 3, 4, F=1 eV, referenced to the Ag/AgCl electrode.

The interaction energy E_{int} for formation of the complex from the subunits is defined as:

$$E_{int} = E_{comp.} - E_A - E_B \tag{3.1}$$

where $E_{comp.}$, E_A, and E_B are the B97-D/Def2-TZVP//B97-D/6-311G(2d,p) energy of the complex, and the subunits A and B, respectively. Two cases were considered for the interaction energy of complex formation. The first considered all species in their fully relaxed state and the interaction energy is referred to as E_{int}^{relax}. The alternative case considers the individual subunits in the configuration adopted within the complex, and the interaction energy is referred to as $E_{int}^{restrict}$. In this regard, the primary difference between E_{int}^{relax} and $E_{int}^{restrict}$ is due to geometry change.

In the discussions of complex electronic rearrangement, it is useful to consider the rearrangement of the total electronic density caused by the interaction between monomers. The resulting electronic change is defined as

$$\Delta\rho = \rho_{comp.} - \rho_A - \rho_B \tag{3.2}$$

where $\rho_{comp.}$ is the electron density of the complex and ρ_A and ρ_B are the electron densities of the subunits forming the complex in exactly the configuration adopted in the relaxed complex. Within this definition, a positive value (depicted in yellow in section A.) indicates electron accumulation, and a negative value (depicted in blue in section A.) indicates electron depletion.

For comparison of geometry in the pentaindenocorannulene systems, it is useful to define the cone angle, α depicted in Fig. 3.1.

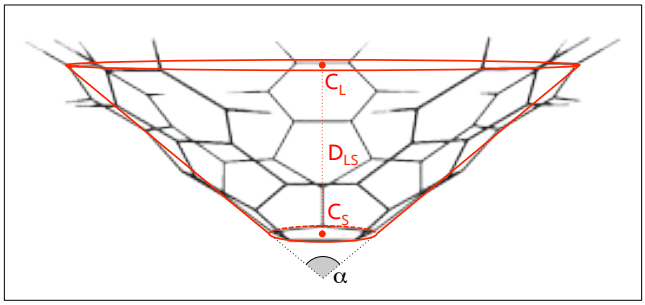


Figure 3.1: Cone angle, α defined for pentaindenocorannulene. The top circle of the cone centered at C_L is defined using the average coordinates of the ten upper carbons. The bottom circle of the cone centered at C_S is defined using the average coordinates of the five lower carbons.

Within this frame, the cone angle is defined as

$$\alpha = 2 \times \tan^{-1}\left(\frac{\mathrm{R} - \mathrm{r}}{\mathrm{D_{LS}}}\right) \tag{3.3}$$

where R is the radius of the circle associated to centroid C_L, and r is the radius of the circle associated to centroid C_S. D_{LS} is the distance between the two centroids C_L and C_S.

Visualization of the total electronic rearrangement was achieved with in-house developed method for computing the density differences and wxMacMolPlt[214] for the rendering. Geometrical parameters, e.g., distances were measured with mercury,[215] and cone angles were obtained with our method.

III Results and discussion

Towards consideration of one complete unit cell of the crystal structure, it was of interest to first carry out calculations on smaller PIC complex derivatives to understand step-by-step the crystal growth. Second the structural results of the aggregation of PIC with C_{60} and C_{70} are presented. The remainder of this section reports the properties of the most stable structures: NMR data, reduction potentials and electronic spectrum.

A. Structural Results

Complexation of pentaindenocorannulene

The structure of pentaindenocorannulene (PIC) is considered as a two-unit complex before involving interaction involved in the trimers are studies. Finally the structures reported by L. T. Scott *et al.*[178;195] and by S. Lampart *et al.*[196] are investigated. Particular consideration is given to the electronic structure and interaction energy along the stacking axis. Comparisons were made between the crystal structure and the B97-D/6-311G(2d,p) optimized structure (Fig. 3.2). The interaction energies for these complexes as obtained with eq. 3.1 are summarized in Table 3.1.

	Structure A		Structure B		Structure C	
	E_{int}^{relax}	$E_{int}^{restrict}$	E_{int}^{relax}	$E_{int}^{restrict}$	E_{int}^{relax}	$E_{int}^{restrict}$
Gas phase	-48.23	-48.50	-42.98	-43.59	-33.62	-34.04
Solvent	-44.68	-45.06	-39.51	-40.27	-29.70	-30.28

Table 3.1: B97-D/Def2-TZVP//B97-D/6-311G(2d,p) calculated interaction energies in kcal/mol. Structures A, B, and C are depicted in Fig. 3.2.

From a structural point of view, the cone angle (see Fig. 3.1, and eq. 3.1) varies between 102.15° (Fig. 3.2 (C1)) and 105.75° (Fig. 3.2 (B1)), which corresponds to a 0.55 kcal/mol energy range. For comparison, the cone angle of pentaindenocorannulene is 103.51° at the same level of theory. It is therefore not surprising that structure A is favored over structures B and C. Indeed, the former dimer presents the smallest structural change between the geometry of the units within the dimer and the geometry of the infinitely separated monomers. The top monomer is 1.54° wider than PIC_1 and the bottom monomer is 0.06° narrower than PIC_1, Ã§Ã§which corresponds to a stability loss of only 0.15 kcal/mol and 0.12 kcal/mol, respectively. Notably, the resulting 0.27 kcal/mol (*i.e.*, 0.12 + 0.15 kcal/mol) difference due to structural change corresponds to the difference between E_{int}^{relax} and $E_{int}^{restrict}$, establishing that the chosen methodology does not suffer from any artifacts such as basis set superposition error (BSSE).

Structure B shows a larger geometry change from monomers to dimer, resulting in a less stable dimer (5.25 kcal/mol less stable than observed in structure A). The lower and upper units forming the complex have cone angles 1.09° and 2.24° wider than in the relaxed pendaindenocorannulene, corresponding to 0.30 and 0.31 kcal/mol difference in energy, respectively. Further in the discussion, a perfect alignment

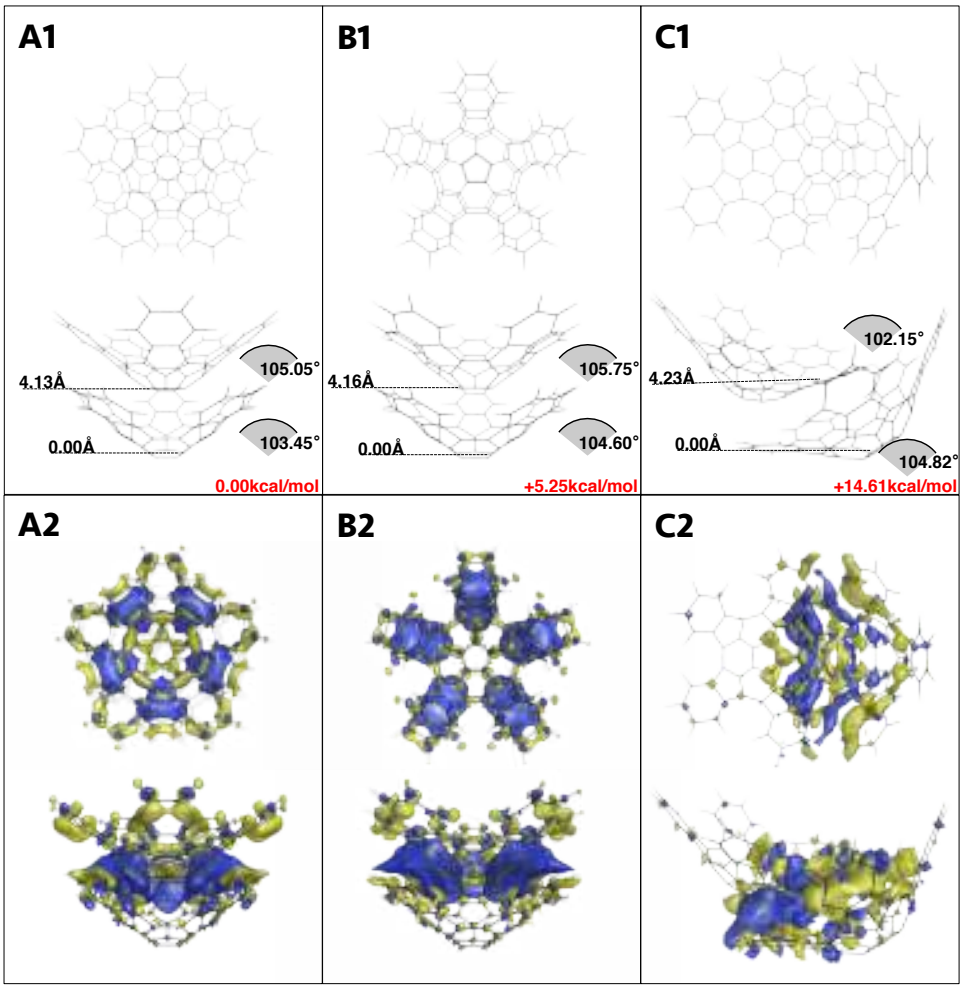


Figure 3.2: (A1, B1, C1) B97-D/6-311G(2d,p) optimized structures in gas phase. Structure C is half the unit cell of the crystal structure published by L. T. Scott *et al.*,[178;195] and structures A and B are two conformers of the new experimental crystal structure.[196] Structure B is the most stable conformation, from both experimental and theoretical perspective. (A2, B2, and C2) illustrate the total electronic density change upon dimerization (see eq. 3.2). Contours in yellow show electron density accumulation and in blue electron density depletion. A contour cutoff value of $0.0002e^{-}/Å^{3}$ was used.

of the stacking units is referred to as a *defect*.

Structure C, which is half of the unit cell of the crystal by L. T. Scott *et al.*, is 14.61 kcal/mol less stable than structure A, although the cost caused by structural change is smaller than that observed in structure B. In structure C, CH-to-π interactions control the packing order.

This preference for structure A, presenting a 36° angle between the individual bowl was also observed experimentally by means of synchrotron radiation. Notably, structure A benefits from a strong dipole moment and large vdW surface.

In the analysis of the electronic structure changes upon dimerization, the actual geometry parameters are seen to only partially explain the preferred conformer. It becomes also insightful to look at the total electronic density redistribution upon dimerization. As a matter of fact, a closer look at Fig. 3.2 (2) reveals that only half of the electronic density of each of the monomers of dimer C appears to participate in the dimerization process. On the other hand, in complexes A and B, the entire electronic

density of both monomers appears to be involved in the dimerization. This reflects larger vdW surface benefit in the case of columnar-like stacking (complex A and B) over the atypical stacking motif of complex C.

Towards consideration of one complete unit cell of the crystal structure, calculations of the three-monomer complex were carried out. In this case, the four possible arrangements of the three subunits along the stacking axis are: (i) the three monomers are aligned, referred to as a A-A-A pattern, (ii) two interacting monomers are aligned and the third presents a 36° rotation along the stacking axis, referred to as a A-A-R, and to (iii) R-A-A, and (iv) each monomer stacks with a 36° rotation, referred to as A-R-A.

Fig. 3.3 illustrates the B97-D/6-311G(2d,p) optimized structures and associated total electronic density redistribution plots for the four cases above-mentioned. Of the four arrangements, structure A appears to be the most stable in terms of complexation energy. In this structure, the monomer units follow an A-R-A pattern, with a 36° rotation along the stacking axis between each monomer. Structures B and C are less stable energetically by 5.26 and 5.80 kcal/mol, respectively. The corresponding patterns for these two complexes are R-A-A (B) and A-A-R (C), respectively. The least stable trimer is structure D, with all monomer units perfectly aligned in a A-A-A pattern. The latter structure is 10.85 kcal/mol less stable than structure A.

A closer look at the cone angles in the dimer and trimer complexes revealed similarities. The lower monomer of dimer A in Fig 3.3 (A) can be compared to the lower units of trimers A and B, and to the middle unit of trimer C (Fig. 3.3) where all four units present a 36° angle with the monomer sitting on top. In all four cases, the cone angle spans a range of 103.45° - 103.59°. Similarly, the lower unit of trimer C has the same cone angle as the lower monomer of dimer B. This similarity in cone angle is also observed when comparing the top monomers of trimers A, B, and C and their corresponding top units in dimers A and B.

Comparing the calculated interaction energies of the trimers with the interaction energies in Table 3.1, one can see that the *ca.* 5 kcal/mol difference resulting from the rotation of a monomer along the stacking axis is independent of (i) the number of monomer units forming the complex, and (ii) the position of the defect along the stack. In addition, trimer C with all monomers aligned, is *ca.* 10 kcal/mol less stable than trimer A, corresponding to twice the cost of the *defect*.

To carry out a more accurate comparison of the calculated vs. two experimentally observed crystal structure forms, an analysis of one complete unit cell with four monomers was considered. In particular, it is known that the two polymorphic crystal structures differ in packing. The unique packing motif presented by L. T. Scott *et al.*,[178;195] described as a stacking of PIC dimers, is very distinctive, not shared with any other curved polycyclic aromatic hydrocarbon (PAH). The new polymorph reported herein however, stacks into infinite bowl-in-bowl columns, similar to the packing motif found in other curved PAHs.[175;176;216;177–179] The bowls face in opposite direction in adjacent columns causing the space group of the crystal to be centrosymmetric (not polar). The PIC molecules in each column are rotated 36° about the column axis with respect to its preceding neighbor leading to a staggered stacking. The same pattern was observed in the dimers of the published polymorph of PIC.[178;195] The packing allows for $\pi - \pi$ interactions between the peripheral six-membered rings of one molecule and the corannulene six-membered rings of the following molecule. The centroid-centroid distances range from 3.560(3) to 3.645(3) Å. This preference for the bowl-in-bowl stacking is also reported in theory

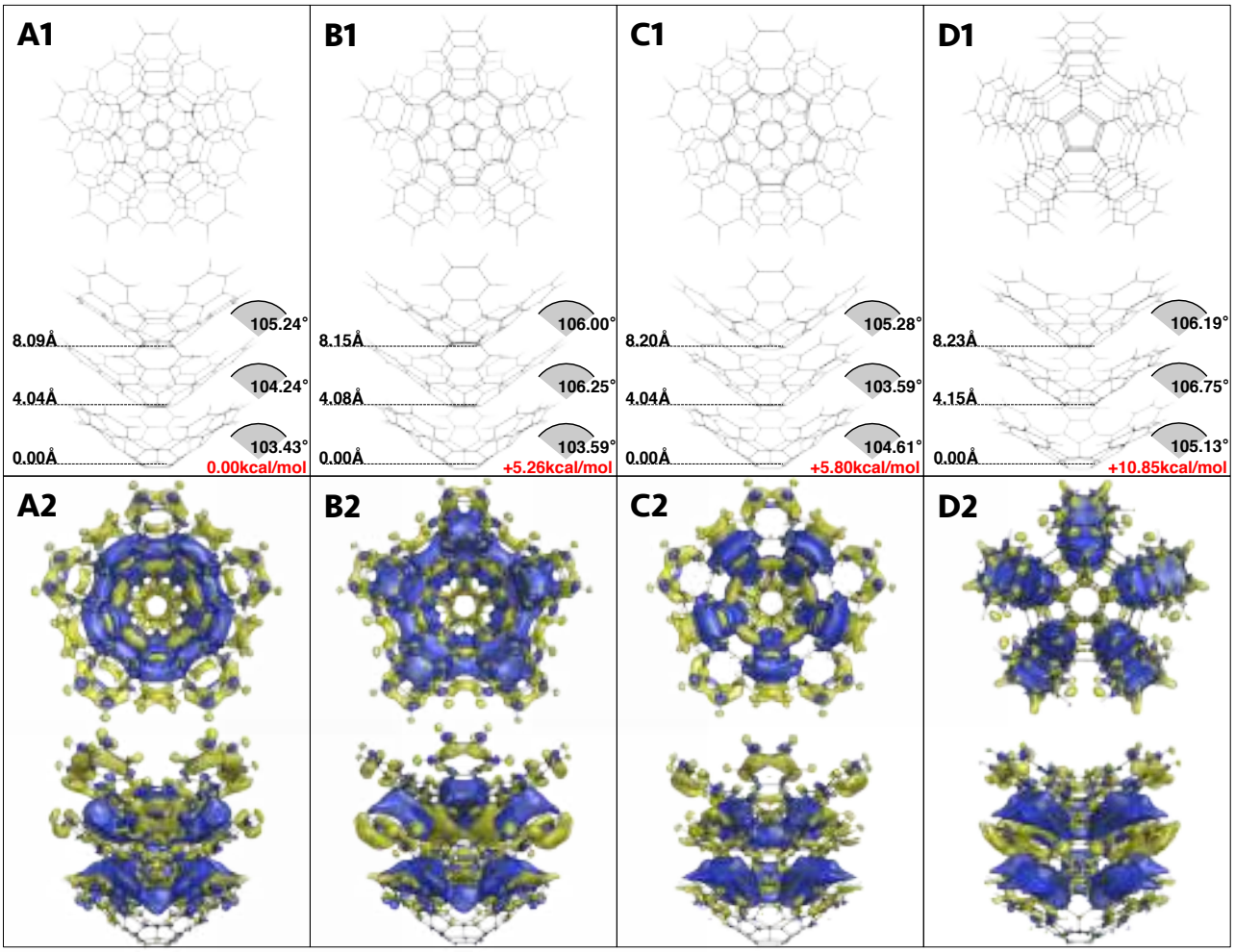


Figure 3.3: (A1, B1, C1, D1) B97-D/6-311G(2d,p) optimized trimers with the four possible arrangements: (A) A-R-A, (B) R-A-A, (C) A-A-R, and (D) A-A-A where A stands for aligned and R for 36° rotation. (A2, B2, C2, D2) Total electronic density redistribution upon trimerization (see Eqn. 2). Contours in yellow show electron accumulation and in blue electron depletion. A contour cutoff value of $0.0002e^{-}/Å^{2}$ was used.

(see Figure 3.4).

The new polymorph (tetramer A) stacking along the b-axis following a A-R-A-R bowl-in-bowl column is by *ca.* 8.5 kcal/mol more stable than the published polymorph (tetramer B). Interestingly, in the staggered dimers, the interaction distances (4.05 and 4.07 Å) are very similar to the ones in the columnar stacking (4.05, 4.03 and 4.04 Å). A close look at the top view of the total electronic rearrangement shows similar pattern in the electronic redistribution. Also, the centroid-centroid distance measured to be 3.95 Å, is by 0.28 Å smaller than the corresponding distance measured in dimer C.

The trend observed with regard to the calculated cone angle in the stacking pattern is confirmed by tetramer A depicted in Fig. 3.4. Of the first and last units of structure A are very close to the calculated α of the monomer units forming dimer A. Furthermore, the two middle units of tetramer A have cone angles very close to the middle monomer of trimer A (*i.e.* A-R-A).

The stability difference between the new polymorph, tetramer A, and the published polymorph (see Fig. 3.4 (B)) is much smaller than between dimer A and half the unit cell of the published polymorph

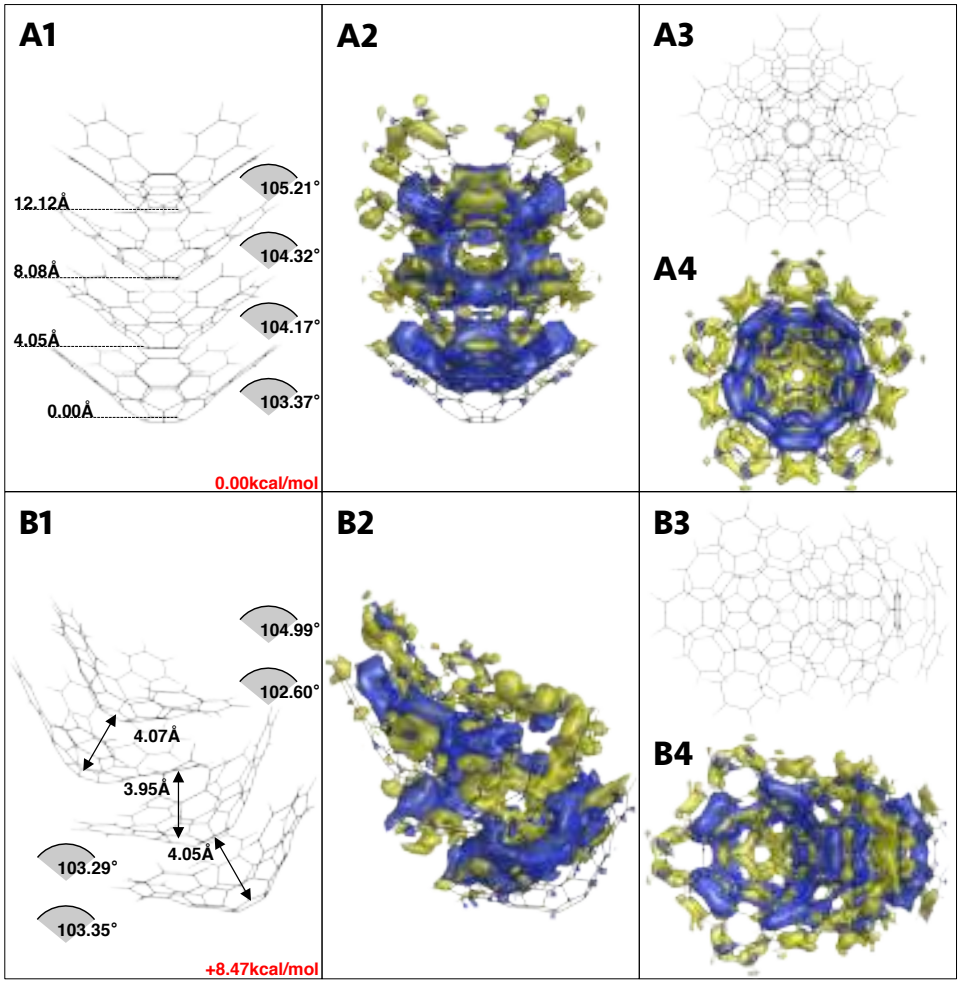


Figure 3.4: B97-D/6-311G(2d,p) optimized tetramers (1) and (3). Recent experimentally determined polymorph structures described in this work (A) and (B) published experimentally determined polymorph structure. (2) and (4) total electronic redistribution upon tetramerization. Contours in yellow show electron accumulation and in blue electron depletion. A contour cutoff value of $0.0002e^{-}/Å^{3}$ was used.

(see Fig. 3.2 (C)). This drop – from 14.61 kcal/mol to 8.47 kcal/mol difference – can be explained by looking at the total electronic rearrangement. The electronic change upon dimerization involves only partial electronic density mixing between monomers, while the electronic redistribution upon tetramerization appears to involve the entire system. Nonetheless, the preference for the columnar-like stacking is reported. Also, such stacking motif benefiting from strong dipole moment and large vdW surface was already reported in literature and is shared with other curved polycyclic aromatic hydrocarbon.[178]

Aggregation with C_{60}

The assembly of polycyclic aromatic hydrocarbons with corannulene-based system was believed to be only present in gas phase until evidence was found for its existence on surface and solid state. Up-to-date, there have been only a handful of examples of corannulene-based complexations with fullerene in solution. The evidence for the aggregation of C_{60} with PIC in solution came by using the method of continuous variations.[217–219] The data* clearly and reproducibly attested an aggregation of C_{60} and

*The ^{1}H NMR chemical shifts of different mixtures of $C_{60}$and PIC with a constant total concentration of 2 mM were plotted against the molar fraction of PIC. The chosen concentration range was well above the limit of PIC self-assembly.

PIC. Furthermore, the obtained maximum of 0.68 in the job plot indicated the possibility of a trimeric assembly (C_{60}@PIC_2) rather than the expected dimer. As was the case for the complexation of PIC, several starting structures were optimized to find the most stable aggregate. Towards consideration of the 2:1 aggregate, we carried out calculations on the 1:1 aggregate to understand the type of interactions involved. Also, for a deeper understanding of concave-convex $\pi - \pi$ interactions, C_{70}@PIC and C_{70}@PIC_2 were investigated in a similar way.

Fig. 3.5 shows the four 1:1 aggregates. A and B are the C_{60}@PIC, and C and D the C_{70}@PIC. Structures A and C have a perfect alignment between PIC and the PIC subunit of C_{60} and C_{70}, respectively, and structures B and D have a 36° rotation along the stacking axis. The similarity between the C_{60}@PIC and the C_{70}@PIC is flagrant, both in term of energetics and in term of structures. In either case, the rotated configuration is favored (see Fig. 3.5 and Table 3.2). Indeed, the perfect alignment leads to a *ca.* 4 kcal/mol destabilization of the aggregates. Interestingly, the cone angle α is narrower when C_{70} is on top of PIC than in the C_{60}@PIC case. Even though the difference is relatively small, α is by 0.11 to 0.18° smaller (Fig. 3.5). A comparison of the interaction distance between aggregates A and C, and between aggregates C and D again show similarity. In the first case, 4.05 Å separated C_{60} from PIC and 4.03 Å for C_{70}@PIC. In the second case, the distance between PIC and C_{60}/C_{70} is 4.19 Å. In line with these observations, the interactions energies are very close.

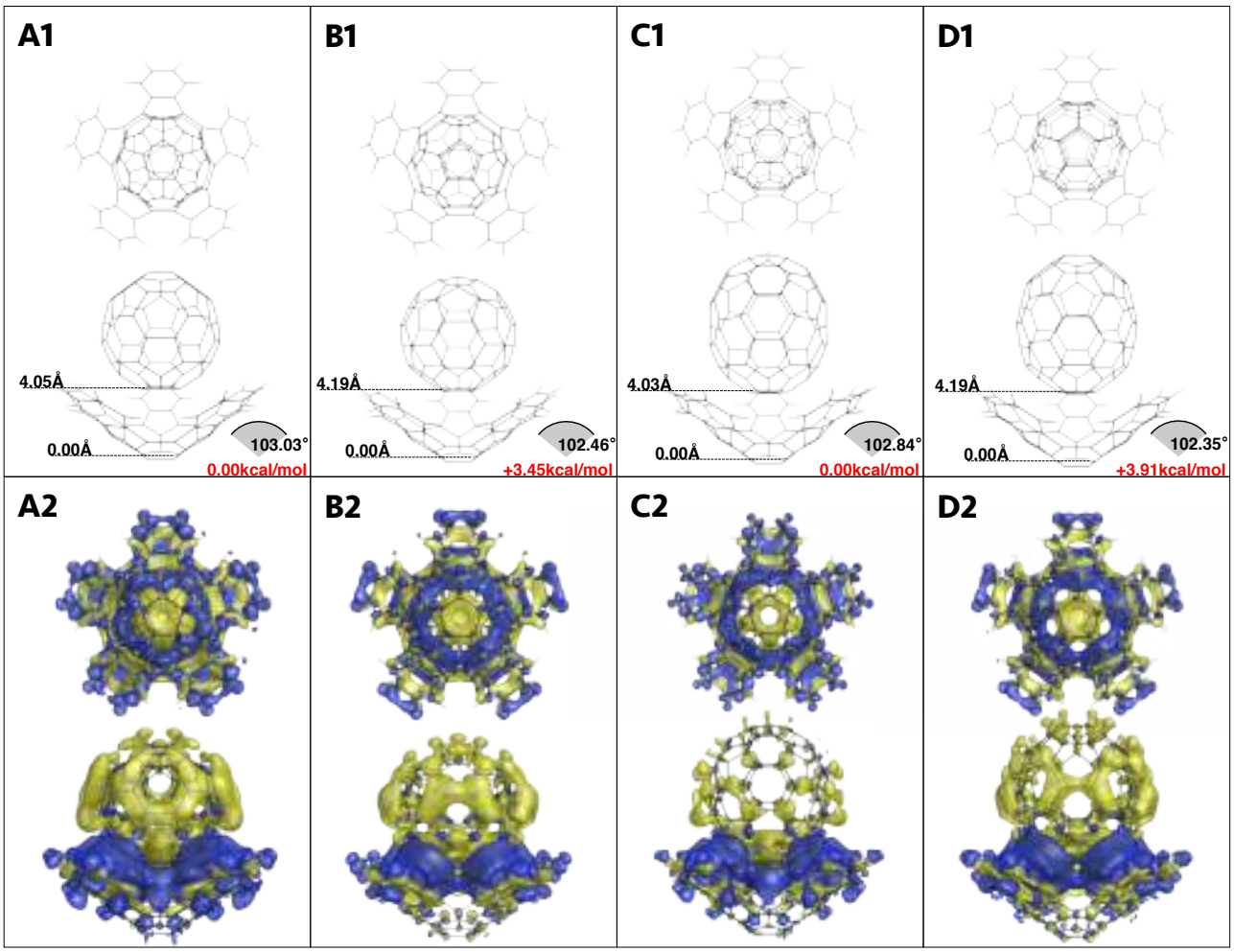


Figure 3.5: B97-D/6-311G(2d,p) optimized 1:1 aggregates. A and B displays C_{60}@PIC and C and D C_{70}@PIC. Structures A and C have a perfect alignment between PIC and the PIC subunit of C_{60} and C_{70}, respectively, and structures B and D have a 36° rotation along the stacking axis

	PIC/PIC/C_{60}		PIC/C_{60}/PIC		PIC/PIC/C_{70}		PIC/C_{70}/PIC	
	E_{int}^{relax}	$E_{int}^{restrict}$	E_{int}^{relax}	$E_{int}^{restrict}$	E_{int}^{relax}	$E_{int}^{restrict}$	E_{int}^{relax}	$E_{int}^{restrict}$
Gas phase	-41.19	-40.84	-37.74	-37.66	-40.87	-41.07	-39.96	-37.42
Solvent	-39.62	-38.58	-36.23	-35.43	-38.43	-38.66	-34.76	-35.23

Table 3.2: B97-D/Def2-TZVP//B97-D/6-311G(2d,p) calculated interaction energies in kcal/mol. Aggregates A, B, C, and D are depicted in Fig. 3.5.

Based on the conclusion drawn with the results obtained on the complexation of PIC along with the 1:1 aggregates, two starting structures per 2:1 aggregate (C_{60}@PIC_2 and C_{70}@PIC_2) were investigated. The four structures are shown in Fig. 3.6 and each of them presents a 36° rotation between each subunit. Aggregates A and B are displays C_{60}@PIC_2, and aggregates C and D C_{70}@PIC_2. In either case, both a *nest* configuration (A and C) and a *sandwich* configuration (B and D) were considered.

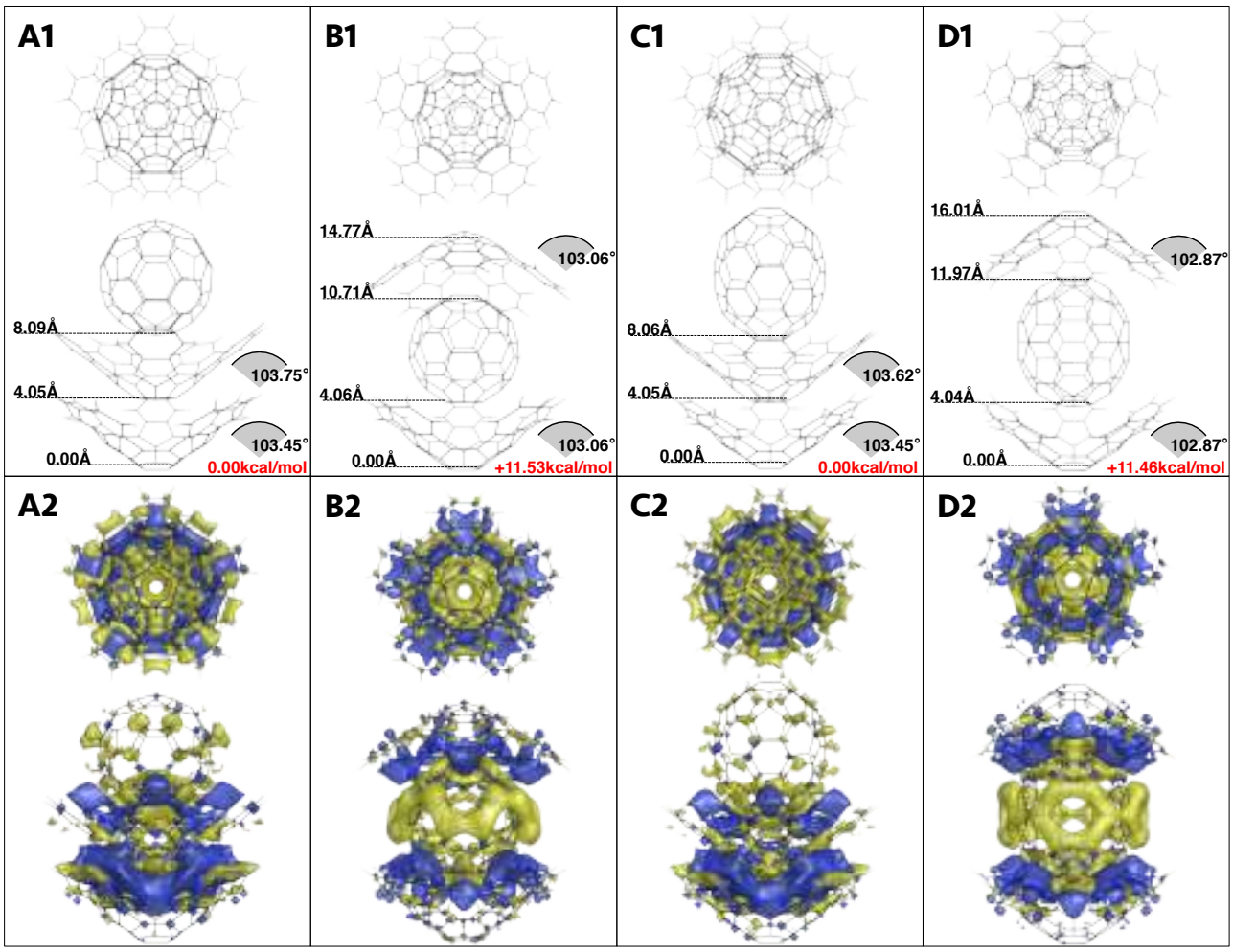


Figure 3.6: B97-D/6-311G(2d,p) optimized 2:1 aggregates. A and B displays C_{60}@PIC_2 and C and D C_{70}@PIC_2. Structures A and C are in a "nest" configuration and structures B and D in a "sandwich" configuration.

The preference for the *nest* configuration is observed in C_{60}@PIC_2 and in C_{70}@PIC_2. As was the case for the 1:1 aggregates, the cone angle α is narrower when C_{70} is involved. This is explained by the the difference in width between C_{60} and C_{70}. Indeed, the distance from the centroid defined by the top carbon atoms of the corannulene subunit of C_{60} and C_{70} and the latter carbon atoms is 3.038 Å and 3.015 Å, respectively. From the 2:1 aggregate C in Fig. 3.6, one can see that the lower PIC unit has

the same cone angle than the lower PIC unit of aggreate A in Fig. 3.6, namely 103.45°. Only the PIC unit directly involved in the interaction with either C_{60} or C_{70} was affected.

Competition between complexation and aggregation

Further structural analysis was performed to understand the competition between the formation of the complex and the formation of the aggregate. Table 3.3 summarizes the interaction energy, $E_{int}^{restrict}$, between a top unit (PIC, C_{60}, C_{70}) and a core unit (PIC, PIC_2, PIC_3). Besides C_{60}@PIC_3 and C_{70}@PIC_3, which are respectively displayed in Fig. 3.7 A and B, all the other structures were introduced in the previous section of this chapter.

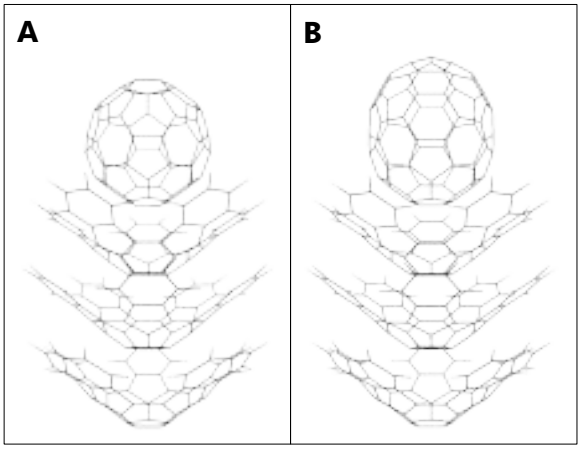


Figure 3.7: B97-D/6-311G(2d,p) optimized 3:1 aggregates. A displays C_{60}@PIC_3 and B C_{70}@PIC_3. Both structures are in a "nest" configuration.

core unit	top unit		
	PIC	C_{60}	C_{70}
PIC	-45.06	-38.58	-38.66
PIC_2	-47.10	-41.23	-41.77
PIC_3	-46.87	-41.52	-41.81

Table 3.3: B97-D/Def2-TZVP//B97-D/6-311G(2d,p) $E_{int}^{restrict}$ of an additional unit (*i.e.* PIC or C_{60} or C_{70}) on top of the main unit PIC, PIC_2 and PIC_3. Interaction energies are in kcal/mol.

From Table 3.3, it is shown that the largest difference in interaction energy is in the two-unit case PIC@PIC ($E_{int}^{restrict} = -45.06$), C_{60}@PIC ($E_{int}^{restrict} = -38.58$) and C_{70}@PIC ($E_{int}^{restrict} = -38.66$). Complexation is favored over aggregation by *ca.* 6.8 kcal/mol. The interaction of the top unit with larger core unit lead to smaller differences in PIC@PIC_2 than in C_{60}@PIC_2. As a matter of fact, the complexation PIC@PIC_2 is favored by *ca.* 5.8 kcal/mol over C_{60}@PIC_2, and by *ca.* 5.3 kcal/mol for PIC@PIC_2 over C_{60}@PIC_3 (similar trend is observed for systems involving C_{70}). Similar findings were already reported in literature.[220;221] However, even though the solvation effects are described via COSab, explicit solvation was not included. Therefore, for a fair comparison, structures involving PIC as top unit should host an ortho-dichlorobenzene solvent molecule in a *nest* configuration. Such structures were fully relaxed in their corresponding Cs symmetry point group. The interaction energies $E_{int}^{restrict}$ between the top unit formed by either solvent@PIC or C_{60}@PIC, and the underlying unit PIC are compared and summarized in Table 3.4.

It was found that the interaction is slightly stronger between the lower PIC and C_{60}@PIC than the lower PIC and solvent@PIC: -47.61 kcal/mol and -44.98 kcal/mol, respectively. It shows that explicit

core unit	underlying PIC
solvent@PIC	-44.98
C_{60}@PIC	-47.61

Table 3.4: B97-D/Def2-TZVP//B97-D/6-311G(2d,p) $E_{int}^{restrict}$ of an underlying PIC unit at solvent@PIC and at C_{60}@PIC. Energies in kcal/mol.

solvation would be required for a (very) accurate competition study between aggregation and complexation. Also, with this simple model, it appeared that, as soon as PIC aggregates with C60, PIC preferably binds to this 1:1 aggregate than demerize.

B. NMR data

Absolute ^{1}H chemical shielding tensors were predicted by means of the CGST[212;213] computational NMR method. For comparison with experiments, the δ_{CGST} was correlated to the conventional δ(TMS) values. Assuming a linear correlation between theory and experiments, one has the relation: $\delta = m \times \delta + C$.[222;223] In the case where the slope is assumed to be $m = 1$, the chemical shift of a nucleus of interest, δ^i, is obtained by computing the CGST chemical shift for TMS, δ_{CGST}(TMS), and taking the difference between that value and the absolute shift computed for the nucleus of interest, δ^i_{CGST}. This leads to an approach referred to as the *shifted method*:

$$\delta^i_{shifted} = \delta_{CSGT}(\text{TMS}) - \delta^i_{CSGT} \tag{3.4}$$

Although the assumption of linearity appears to be well founded, the slope of the line is in general not unity and depends on the computational method and basis set.[223] This leads to an approach referred to as the *correlated method*:

$$\delta^i_{corr.} = m \times \delta^i_{CSGT} + C \tag{3.5}$$

where m is the slope of the correlation line, and C the intercept. In order to obtain useful chemical shift predictions over the full spectral window, these empirical correlations must display a high degree of linearity ($R^2 \geq 0.999$). Fig. 3.8 shows the reference set of systems to establish the required correlation line. We decided to use a (relatively) small representative set of structures that nonetheless spans a reasonable range of chemical shifts. The compounds chosen were TMS δ(0.00), 1,1,2,2,-tetrachloroethane-D2 δ(5.98), benzene δ(7.38), toluene δ(2.36, 7.19, 7.27), and pyrene δ(8.03, 8.10, 8.20). The ^{1}H shifts were measured in TCE in this group. The resulting calibration formula (eq. 3.5) was determined to be: $\delta^i_{corr.} = -1.0034 \times \delta^i_{CSGT} + 31.454$ (goodness of the fit: R2=0.99899). The five systems were optimized at the B97-D/Def2-QZVPPD level of theory in the gas phase. Table 3.5 compares the experimental with the computed chemical shifts of the reference set (Fig. 3.8).

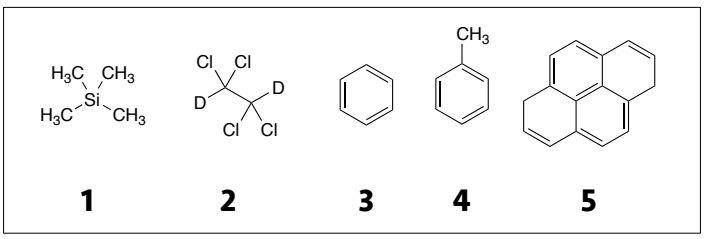


Figure 3.8: Reference set used to establish the correlation line. It is composed by TMS (1), 1,1,2,2-tetrachloroethane-D2 (2), benzene (3), toluene (4), and pyrene (5).

Compound	Exp.	B97-D/cc-pVTZ			B97-D/Def2-TZVP		
		δ_{CSGT}	$\delta_{shifted}$	$\delta_{corr.}$	δ_{CSGT}	$\delta_{shifted}$	$\delta_{corr.}$
(1) TMS	**0.00**	31.36	0.00	0.07	31.29	0.00	0.06
(2) TCE	**5.98**	25.68	5.68	5.70	25.56	5.73	5.81
(3) benzene	**7.38**	23.99	7.37	7.37	24.04	7.25	7.34
(4) toluene	**2.36**	29.02	2.34	2.39	28.99	2.29	2.36
	7.19	24.17	7.19	7.19	24.14	7.14	7.23
	7.27	24.12	7.24	7.24	24.20	7.08	7.17
(5) pyrene	**8.03**	23.27	8.09	8.08	23.32	7.97	8.06
	8.10	23.13	8.23	8.22	23.14	8.14	8.23
	8.20	23.09	8.26	8.26	23.13	8.16	8.25
RMSD	–	–	0.113	0.109	–	0.120	0.087

Table 3.5: Computed B97-D/cc-pVTZ//B97-D/Def2-QZVPPD, B97-D/Def2-TZVP//B97-D/Def2-QZVPPD and experimental δ(TMS) chemical shifts for the reference set depicted in Fig. 3.8.

Inspection of the correlation parameters show that B97-D/Def2-TZVP yields a slope closer to 1 (m = 1.0034) than B97-D/cc-pVTZ (m = 0.9903). Both basis sets resulted in what would appear as reasonable linear fits: R^2 = 0.99841 and R^2 = 0.99899, for cc-pVTZ and for Def2-TZVP, respectively. The results obtained with the Ahlrichs-style basis set diplays a slightly better correlation between CGST and experimental chemical shifts, tipping the balance in its favor over the Dunning-style basis set. In addition, statistical analysis support this preference. The RMSDs reported in Table 3.5 reflects the better performance of the B97-D/Def2-TZVP *correlation method* over other techniques. This preference for B97-D/Def2-TZVP is supported by the ^{1}H chemical shift obtained for corannulene: at the B97-D/Def2-TZVP level, δ(TMS)=7.89, and at the B97-D/cc-pVTZ δ(TMS)=7.92 using the *correlation method* (experimental δ(TMS) is 7.80[223]).

Fig. 3.9 and 3.10 highlight the symmetry unique hydrogen atoms. All systems belonging to the C_{5v} symmetry point group, each subunit counts two sets of identical hydrogen atoms, depicted with the same color (*i.e.* blue, cyan, red, orange, green, lime, gray and silver).

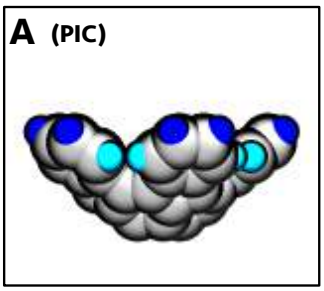


Figure 3.9: PIC structure used for the prediction of the ^{1}NMR chemical shifts in TCE. The two sets of identical hydrogen atoms are depicted in blue and in cyan.

The single PIC unit (Fig. 3.9) was used as reference value to discuss chemical shifts upon complexation and aggregation. The computed ^{1}H δ(TMS), 7.39 and 8.10 ppm, (see Table 3.6) are in good agreement with the measured chemical shifts at low concentration (2 μM, in Fig. 3.11).

Table 3.6 summarizes the ^{1}H NMR chemical shifts of the PIC monomer (3.9), the complexes (A, C, E in Fig. 3.10) and the aggregates (B, D, F in Fig. 3.10). The color scheme of the proton type refers to

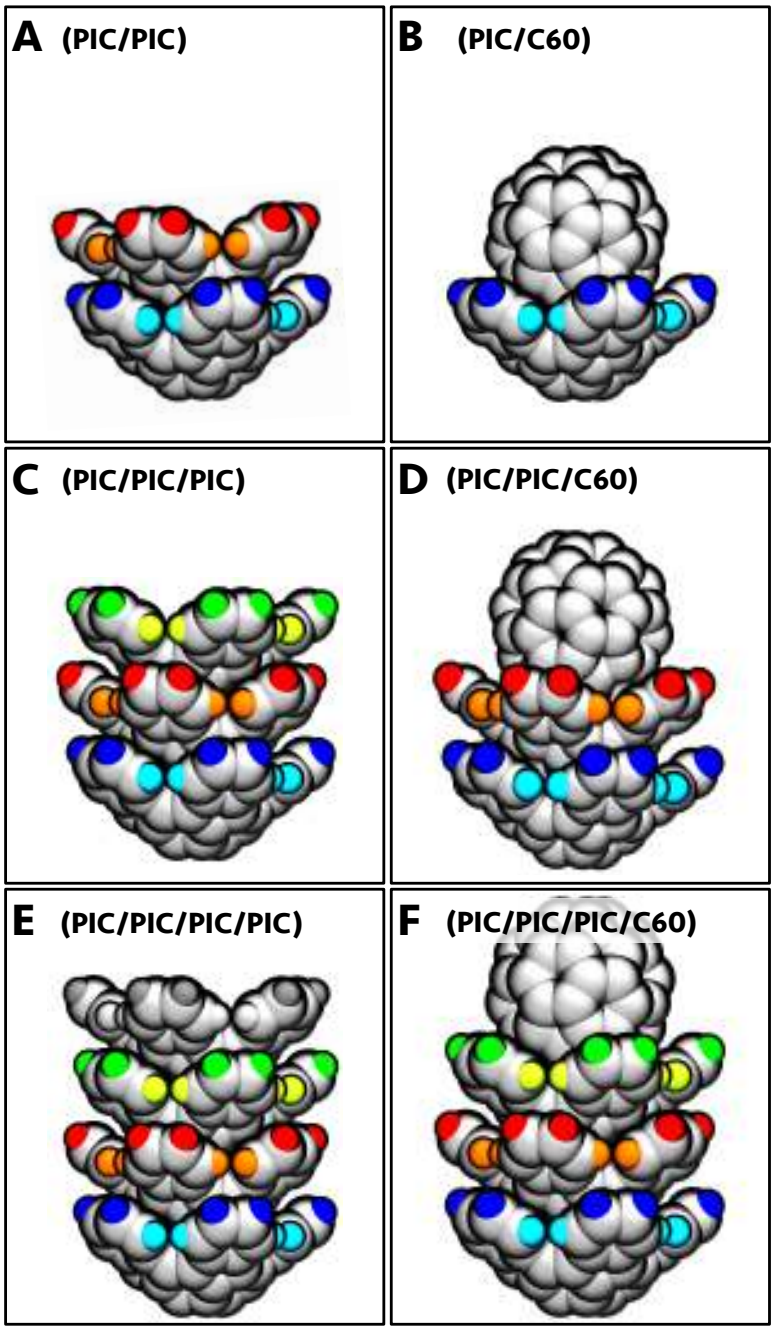


Figure 3.10: Complexes and aggregates used for the prediction of the ^{1}NMR chemical shifts in TCE. The sets of identical hydrogen atoms are depicted in blue, cyan, red, orange, green, lime, gray and silver.

the color scheme displayed in Fig. 3.9 and 3.10.

The computed ^{1}H chemical shielding tensors in TCE for monomer PIC are in accordance with the experimental observation at low concentration. The shifts of the two C_{5v} symmetry unique H's in PIC are 7.39 ppm (in cyan, experimental: 7.43) and 8.10 ppm (in blue, experimental: 8.13 ppm). An increase in the complexation index leads to a shielding effect. As a matter a close look at the blue and cyan proton chemical shifts show this shielding effect, particularly pronounced for the inner proton (in cyan). This is in line with the experimental observation (see Fig. 3.11) and with previous studies.[224;195] The δ(TMS) of both blue and cyan protons appeared to converge to 7.20 and 7.57 ppm, respectively. Similar trend was observed with the red and orange protons. Interestingly, in the PIC/PIC/PIC/PIC

Systems	Proton type							
	●	●	●	●	●	●	●	●
PIC	7.39	8.10	-	-	-	-	-	-
PIC/PIC	7.34	7.65	7.21	7.72	-	-	-	-
PIC/PIC/PIC	7.24	7.59	7.14	7.25	7.08	7.51	-	-
PIC/PIC/PIC/PIC	7.21	7.57	7.03	7.20	7.00	7.04	7.01	7.41
PIC/C60	7.52	8.29	-	-	-	-	-	-
PIC/PIC/C60	7.36	7.70	7.19	7.68	-	-	-	-
PIC/PIC/PIC/C60	7.26	7.62	7.16	7.31	7.06	7.46	-	-

Table 3.6: Chemical shifts (δppm) at the B97-D/cc-pVTZ//B97-D/6-311G(2d,p) level in TCE.

case, the two sandwiched PICs have very similar δ(TMS), the outer proton in particular (red and green, respectively at 7.03 and 7.00 ppm). Finally, it can be seen that the difference between the outer and inner protons, referred to as $\Delta\delta$(TMS) of the top unit (blue vs. cyan, red vs. orange, green vs. lime, and gray. vs. silver) diminished as the number of underlying units increased: $\Delta\delta$(TMS) = 0.71, 0.51, 0.43, and 0.40 ppm. The trend converges to $\Delta\delta$(TMS) = 0.38 ppm.

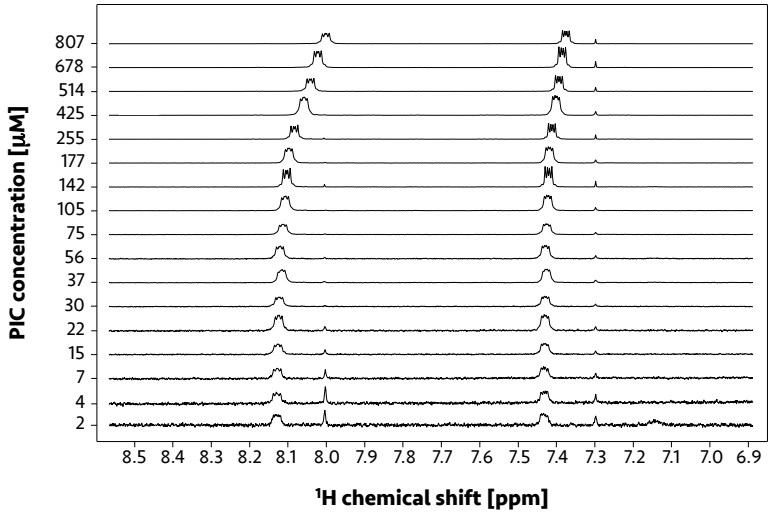


Figure 3.11: Measured ^{1}H NMR chemical shifts in TCE.

Very similar observations and conclusions can be drawn from the results obtained on the aggregates. However, emphasis is placed on the fact that C_{60} appeared not to influence the ^{1}H NMR of PIC. The δ(TMS) of C_{60}@PIC_2 were within 0.05 ppm difference of the PIC/PIC δ(TMS). The comparison of C_{60}@PIC_2 with PIC/PIC/PIC supports this observation, indicating that C_{60} has no important role in the shielding of ^{1}H NMR chemical shift of PIC. Consequently, it is suggested that NMR cannot be used to either prove or contradict the existence of PIC:C_{60} aggregates.

C. Reduction potential

The reduction potential was determined with eq. 3.6 at the B97-3/Def2-TZVPD(THF)//B97D/6-311G(2d,p) level.

$$E^{\circ}_{\mathrm{PIC/PIC}^{n-}} = \frac{\Delta E_{B97-3}}{n \times F} - \left(E^{\mathrm{Fc^+/Fc}}_{\mathrm{SHE}} - E^{\mathrm{Ag/AgCl}}_{\mathrm{Fc^+/Fc}}\right) \tag{3.6}$$

where F is the Faraday constant, n the charge, $E^{\circ}_{\mathrm{PIC/PIC}^{n-}}$ is the reduction potential of PIC/PIC^{n-}, ΔE_{B97-3} is the energy difference between uncharged PIC and the anion PIC^{n-}, $E^{\mathrm{Fc^+/Fc}}_{\mathrm{SHE}}$ is the reference Fc$^+$/Fc redox potential vs. the SHE (4.98 V) and $E^{\mathrm{Ag/AgCl}}_{\mathrm{Fc^+/Fc}}$ the reference Ag/AgCl redox potential vs. the Fc$^+$/Fc (0.64 V). The rationale for using the B97-3 hybrid functional was non-local exchange, which, in this particular case, was required to converge the anionic wavefunctions.

The extended π-surface of PIC allowed the recording of four of its anionic oxidation states by cyclic voltammetry (see Fig. 3.12), which was also observed in a previously reported π-extended bowl-sheet hybrid corannulene.[225;226] However, since the anionic peak potential of the fourth peak was not distinct enough, only three reduction potentials were experimentally quantified: -1.58, -1.96 and -2.45 V. The four anionic oxidation states of PIC were predicted by B97-3/Def2-TZVPD level of theory. The values referenced to the Ag/AgCl electrode (see eq. 3.6) are -1.56, -2.03, -2.39 and -2.80 V.

Consistent with expectations based on previous studies of compounds with extended π-surfaces and increased bowl-curvature,* the first reduction potential of PIC was lower by *ca.* 0.6 V (in THF) compared to mono-indenocorannulene, and by ca. 1 V compared to corannulene, due to a curvature-stabilized LUMO. The latter reduction potential was measured to be -2.5 V and computed to be -2.51 V.

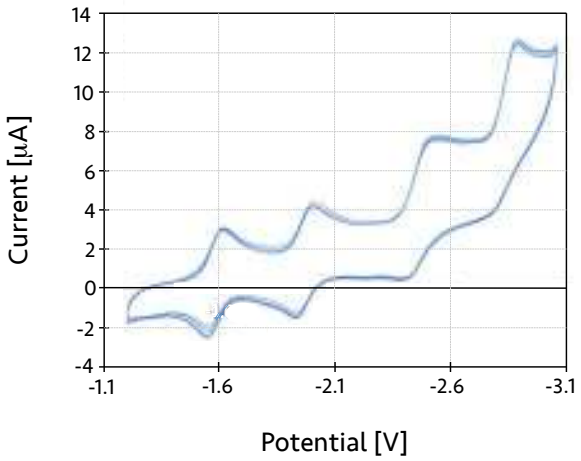


Figure 3.12: Measured reduction potentials of PIC measured in THF and corrected to Fc / Fc+ (+ 0.085 V).

D. TD-DFT spectrum

The TD-DFT spectra were computed at several level of theory using CAM-B3LYP exchange-correlation functional. The performance of CAM-B3LYP was demonstrated in several benchmark studies.[228–233] The simulated spectra were obtained from the oscillator strengths by adding Gaussian line shapes, following the Harada-Nakanishi equation.[234] The extinction coefficients ϵ reads:

$$\epsilon(\nu) = \frac{f_i}{3.483 \times 10^{-5} \times \sqrt{\pi} \times \sigma} \times exp\left[-\left({}^{(\nu-\nu_i)}/_{\sigma}\right)^2\right] \tag{3.7}$$

where ν is the excitation energy in eV and σ is a parameter, chosen to equal 0.075 eV, as small values of σ allow well-defined excitation bands.

*Corannulene is known to have four anionic oxidation states, which were documented electrochemically up to the third reduction.[227;226]

Basis set and solvation dependence

The impact of the basis set choice and solvation effects on the transition energies of PIC and C_{60} were studied by investigating three basis sets – 6-31G(d), and 6-311G(2d,p) from J. A. Pople and Def2-TZVP from R. Ahlrichs – in both gas phase and solution (chloroform). Fig. 3.13 and 3.14 display the spectra of PIC and C_{60}, respectively. In both figures, the results obtained with the 6-31G(d) basis set are depicted in yellow, 6-311G(2d,p) in red, and Def2-TZVP in blue. The continuous lines show the results in solvent and the dashed lines in the gas phase.

Fig. 3.13 and Fig. 3.14 exhibit a red shift when increasing the basis set size and upon inclusion of solvation effects. As a matter of fact, by doubling the number of basis functions from 6-31G(d) to 6-311G(2d,p) a red shift was observed both in the gas phase and in chloroform. Similarly, from the split valence triple-ζ 6-311G(2d,p) set to the full triple-ζ Def2-TZVP basis set yielded a red shift, yet smaller. Upon inclusion of the solvation effects, not only a red shift was observed, but also a increased transition intensity. Nonetheless, the general features were generally well described at either level of theory investigated herein. For the sake of consistency, the Def2-TZVP basis set was used to study the impact of complexation and aggreation on the TD-DFT spectrum.

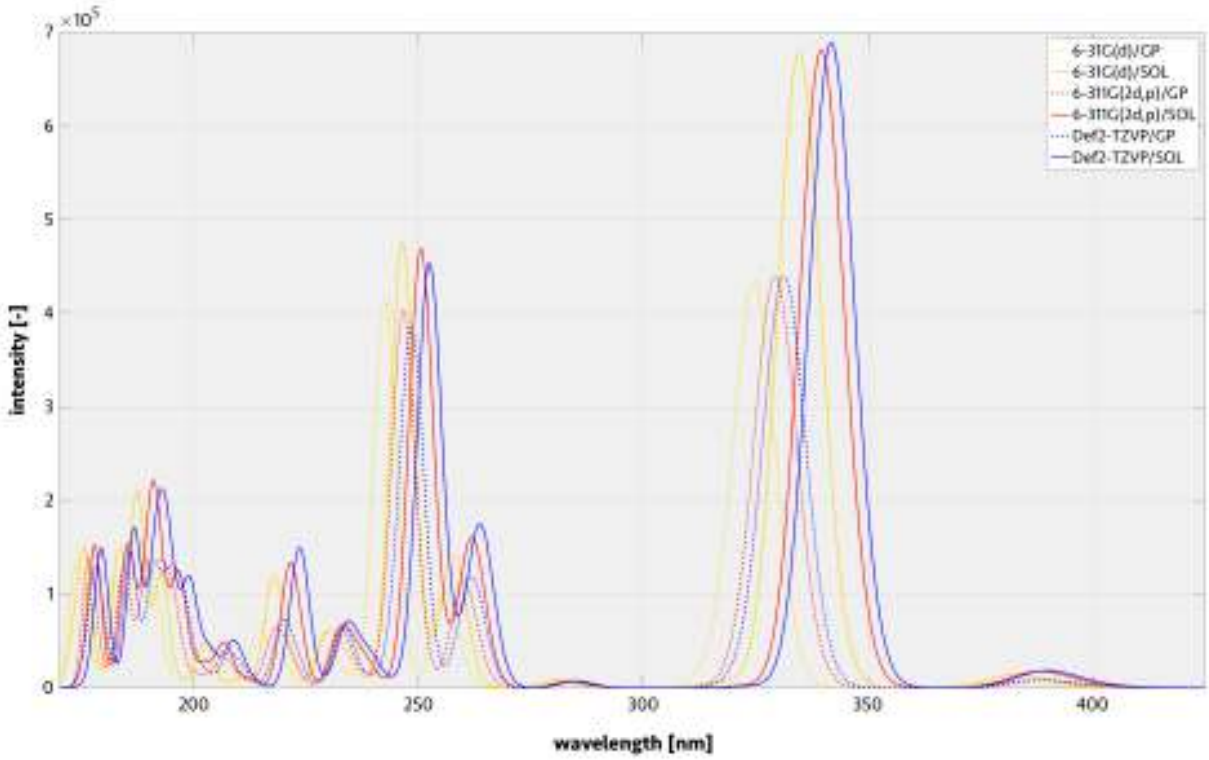


Figure 3.13: TD-CAMB3LYP specta of PIC. In yellow the 6-31G(d) basis set, in red 6-311G(2d,p) and in blue Def2-TZVP. Continuous lines show results in solvent (SOL) and dashed lines in the gas phase (GP).

TD-DFT spectrum of the complexes and aggregates

The results of the TD-DFT calculations are shown in Fig. 3.15.

A close look at the complexation (plain lines in Fig. 3.15) reveals that increasing the number of interacting PIC units did not lead to drastic spectrum changes. However, a red shift was noticed at low wavelengths (175-225 nm). The two intense peaks at 250 nm and 340 nm increased in intensity

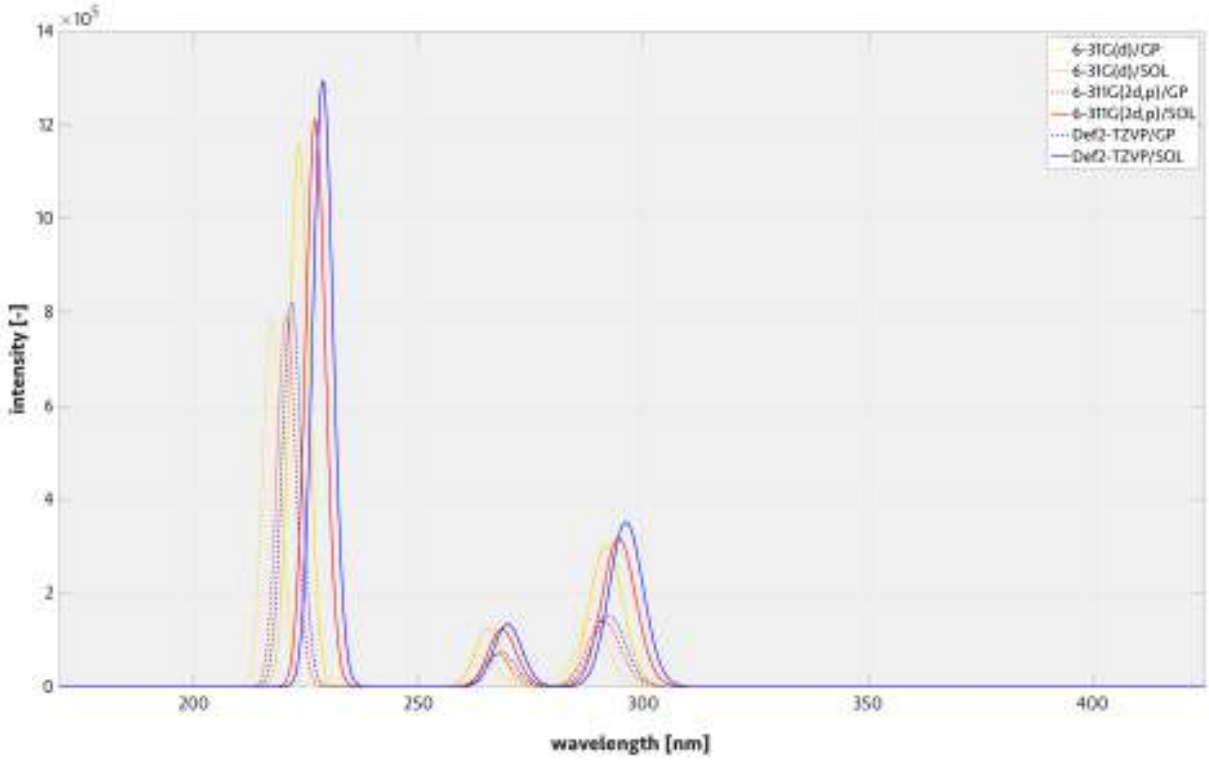


Figure 3.14: TD-CAMB3LYP specta of C_{60}. In yellow the 6-31G(d) basis set, in red 6-311G(2d,p) and in blue Def2-TZVP. Continuous lines show results in solvent (SOL) and dashed lines in the gas phase (GP).

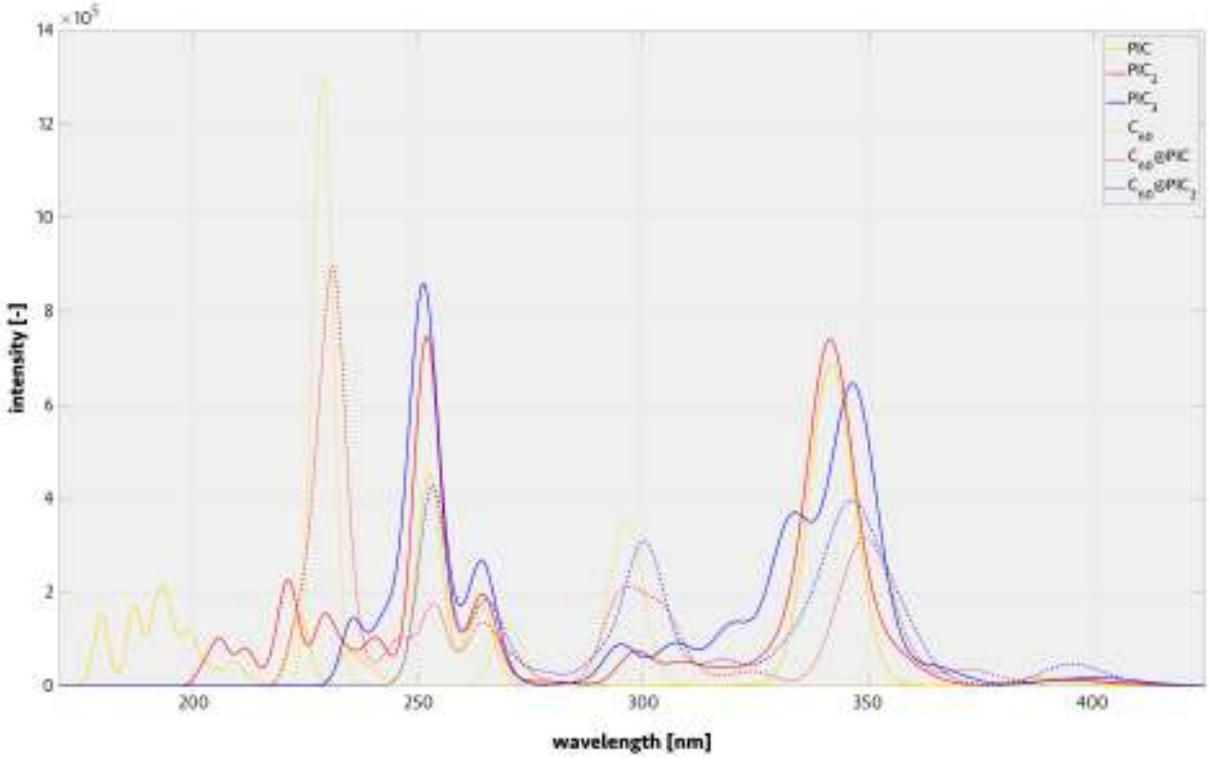


Figure 3.15: TD-CAMB3LYP specta with the Def2-TZVP basis in solvent. In yellow the single-unit system, in red the two-unit systems and in blue the three-unit systems. Continuous lines show the results for the complexation of PIC, and the dashed lines the aggregation of PIC with C_{60}.

with increasing the number of subunits. The peaks between 295 and 335 nm which are not present on the single unit system appeared with dimerization and their intensities even increased on the 3-subunit system. Results for the aggregates (dotted lines in Fig. 3.15) showed a threshold at the 2:1 aggregate. Indeed, the intense peak at 225 nm for C_{60} slightly decreased in intensity upon aggregation with PIC and disappeared in the C_{60}@PIC_2 case. Also, the two peaks observed in the complexation at 250 and 265 nm appeared when investigating C_{60}@PIC_2, while C_{60} alone did not have such transitions. The fact that excitation spectra cannot be used to determine the presence or absence of complexes and/or aggregates was also found experimentally.

IV Conclusion

The study of pentaindenocorannulene and its ability to complex and to aggregate is an illustration of synergistic studies between theory and experiments. In particular, we have demonstrated that the structures and their relative stability agree with crystallographic results. The preference for columnar-like stacking is preferred over the stacking of PIC dimers. In the former case, the columnar stacking benefits from a strong dipole moment and large vdW surface, while CH-to-π interactions dominate and control the packing order in the latter case. We found that the reduction potentials were in accordance with measurements, enabling the characterization of the fourth anionic states.

Both computed and measured NMR confirmed the self-assembly by a deshileding effect upon complexation. However, UV-vis spectroscopy and TDDFT calculations did not show new absorption bands for the aggregate .

Chapter 4

Highly available HPC system for reliable quantum chemistry simulations

I Introduction

The race for scientific discovery by running applications on the fastest machines available for a significant amount of time (*i.e.* weeks and months), while demanding high throughput without interruption, has forced a re-design of high-performance computing (HPC) infrastructures. In this regard, such infrastructures must be able to run in the event of frequent failures in a way that the performance, accessibility and availability is not severely degraded.[20;21]

The introduction of the *Beowulf* cluster[21] systems in the late 90's could be considered as the starting point of parallel computing. This rather simple architecture, made of commodity off-the-shelf components, has been proven to be very efficient.[21] Furthermore, it allows on-demand customization, full flexibility and high performance-cost ratio: three highly desirable characteristics. Although simple, the *Beowulf* cluster met the basis requirements for an HPC system: made of nodes which communicate over a network. In addition, each node contains one or more processors, disks and memory shared by all processors either within the node or over the network. The latter network allows processes and/or information to be shared between the nodes. Nowadays, however, such an infrastructure is outdated and typically requires improved scalability and isolation in case of failures.

The project described in this chapter aims at designing, developing and implementing a highly available HPC for large parallel quantum chemistry calculations and it focuses on efficient redundancy strategies on the hard- and on the soft-ware levels.

Redundancy is a fault tolerance technique, which minimizes the overall failure by introducing the notion of replica. If a service fails, a replica takes over its execution. Defining a level of redundancy is a strategic question when planning a new data center since it has a direct impact on the entire design of the building as well as on the construction and operational costs. It also affects how to integrate future extension plans into the design.[23] The downside of redundancy is that extra resources are required and there is an additional overhead on communication and synchronization.[24] However, via partitioning of the networks such overhead is drastically decreased and most importantly does not affect the production network. In general, large-scale HPC systems[22] may be partitioned, separate interconnected networks may exist to minimize interference, user data and authentication services may be mirrored, to, thus, maximize the overall reliability. Such balancing of the services running on the nodes and such partitioning of the networks, also referred to as high-availability, were strategies used in the design of Arran, the highly available HPC system described in this chapter.

The large number of *ab initio* quantum chemistry packages that needed to be available, lead to specific and complex functional and non-functional requirements. Consequently, the reader has to bear in mind that as each part is tributary of other parts, and, thus, that cross-referencing within the chapter was a necessity.

The remainder of this chapter will focus first on describing the hardware before presenting their (inter)connections in details. For the sake of clarity, the acronyms used throughout the chapter are listed and defined in section II. Sections III and IV glances over the general layout of the new data center, Arran and its hardware specifications. In section V the power redundancy is explained. Then, in section VI and VII, the storage setup and networks architecture are detailed. Sections VIII and X address the configuration of the resources manager and the importance of health checks. Before the concluding remarks, the tools used for mass deployment are carefully described in section IX.

This very challenging project started early summer 2014. It was done jointly with Tyanko B. Aleksiev within the framework of his Master studies in the Department of Mathematics and Computer Science of the University of Udine, in Italy. Figs. 4.2, 4.4, 4.5, 4.6 and 4.7 are taken and adapted from his master thesis.[235]

II List of acronyms

ACRONYMS	MEANING & DEFINITION
▷ arran1	master head-node
▷ arran2	slave head-node
▷ arran	hostname resolved by the DNS, associated to a virtual IP. It is th login point for the users.
▷ arranmngt1	master management-node
▷ arranmngt2	slave management-node
▷ arranmngt	hostname resolved by the DNS, associated to a virtual IP
▷ DRBD	data mirroring software
▷ Pacemaker	HA resource manager used for the migration between master and slave
▷ HPN	high performance network, also referred to as production network: it is used for the calculations
▷ mngt network	management network used for installation, deployment, and updates of the cluster
▷ SPN	service processor network
▷ SLURM	simple Linux utility resource manager, the work-load manager installed on Arran
▷ partitions	used to virtually split the cluster into separate parts to meet individual requirements
▷ accounts	used to enforce restrictions of resources on a set of users
▷ associations	entity used to group information consisting of three parameters: account, partition and user
▷ FAI	fully automatic installation: tool used for mass deployment
▷ Ansible	tools used for deploying the configuration

III Data center

The new data center (DC) of the School of Pharmaceutical Science and Technology (SPST) was designed from scratch to host the new infrastructure. The center accommodates four different servers dedicated to the different fields of application required by SPST.

- ▷ Arran, a 5'568-hyperthreaded core Linux cluster (Ubuntu server 14.04 LTS) shared between three groups
- ▷ Dalmore, a 640-hyperthreaded core Linux cluster (CentOS 6.5) owned by a single group
- ▷ Oban, a Linux web server (Ubuntu server 14.04 LTS) for SPST, which also hosts a virtual host (Windows Server 2012R2)

▷ Dalwhinnie, a Linux server (Ubuntu server 14.04 LTS) providing backup for the whole infrastructure

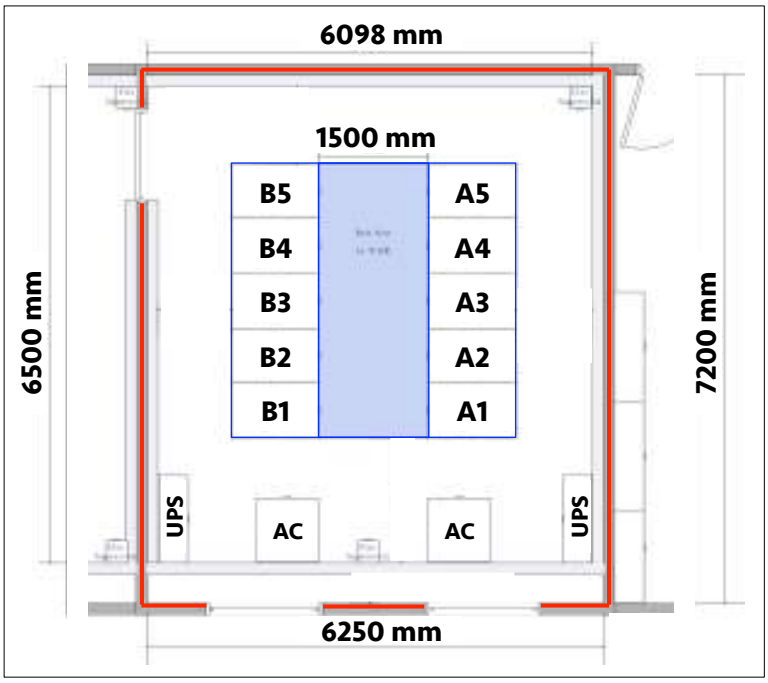


Figure 4.1: Plan of the new data center of the School of Pharmaceutical Science and Technology at Tianjin University. Racks A1-A5 host Arran. Rack B5 hosts Dalmore, Oban and Dalwhinnie. Racks B1-B5 will be used for the future extension of Arran.

IV Hardware and layout

The following subsections outline the technical specifications of the different hardware components in Arran. The components were purchased from DELL (servers and enclosures), DELL FORCE10 (switches) and APC (PDUs and racks).

A. Servers

The rack mounted servers (*i.e.* head-nodes, management-nodes, and compute-nodes) are PowerEdge R630 rack servers, a compact 1U two-socket chassis. The 120 servers were hyperthreaded and Ubuntu 14.04 LTS server distribution was installed on every host.

head-node. The two head-nodes (see Fig. 4.2) – arran1 and arran2 – serve as log-in point for users. It is attached to the `/home/` enclosure storing the users data (see section B. for more details) and runs the main operational services of the cluster. Connection to Arran is granted through the Secure Shell (SSH), a command-line interface and protocol. Authentication is done through public and private keys, a key pair enforcing the connection from verified and secure places where the private key is available.

- ▷ 2× Intel Xeon E5-2690 v3 (12C, 24T)
- ▷ 4× 10K-RPM 2.5” 1.2TB SAS HDD
- ▷ 8× 8GB R-DIMM RAM
- ▷ 2× 1Gbit Ethernet interfaces
- ▷ 2× 10Gbit Ethernet interfaces
- ▷ PERC H730 RAID controller
- ▷ 2× PSU, 750W

management-node. The two management-nodes (see Fig. 4.2) – arranmngt1 and arranmngt2 – host the configuration files and automated softwares required for (i) mass deployment and installation (ii) managing and orchestrating the configurations and (iii) monitoring the health of the whole infrastructure (see sections IX and X for more details).

- ▷ 2× Intel Xeon E5-2690 v3 (12C, 24T)
- ▷ 4× 10K-RPM 2.5" 1.2TB SAS HDD
- ▷ 8× 8GB R-DIMM RAM
- ▷ 4× 1Gbit Ethernet interfaces
- ▷ PERC H730 RAID controller
- ▷ 2× PSU, 750W

compute-node. The 116 compute-nodes (a block of four compute-nodes is highlighted in Fig. 4.2) provide the computer power.

- ▷ 2× Intel Xeon E5-2690 v3 (12C, 24T)
- ▷ 10× 10K-RPM 2.5" 1.2TB SAS HDD
- ▷ 8× 32GB LR-DIMM RAM
- ▷ 2× 1Gbit Ethernet interfaces
- ▷ 2× 10Gbit interfaces
- ▷ PERC H730 RAID controller
- ▷ 2× PSU, 1100W

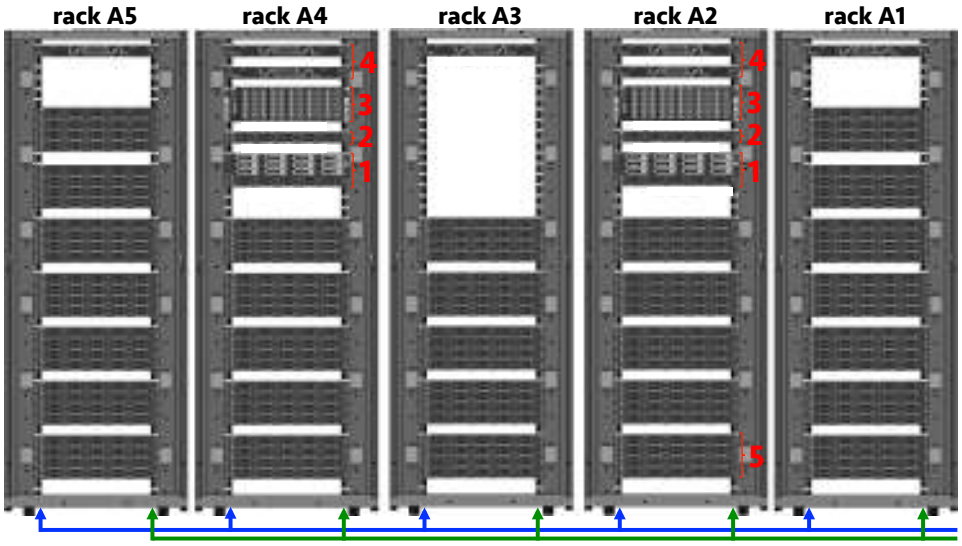


Figure 4.2: Layout of Arran. Is highlighted in (1) the head-node and home enclosure, in (2) the management-node, in (3) the high-performance switches, in (4) the management switches, and in (5) four compute-nodes. The two separate power lines coming from the UPS system are depicted in blue and in green. A1 – A5 refer to Fig. 4.1.

B. Storage

The large amount of data produced by the calculations submitted by users are stored on a PowerVault MD3400 12Gb serial-attached SCSI (SAS) array, a 2U chassis meant for high availability and high performance. Redundancy is provided by an additional physical controller, and by an additional battery which minimizes the risk of data loss in case of an unexpected power cut. The enclosures with the following specifications are depicted in Fig. 4.2.

- ▷ 12G SAS, 2U-12 drive, Dual 8G Cache Controller
- ▷ 12× 4TB 7.2K-RPM Self-Encrypting Near-Line SAS 6Gbps 3.5" HDD

C. Switches

The communication within the cluster goes through two types of switches: Dell Force10 S55 and Dell Force10 Z9500. While the former is used for the management and monitoring network, the latter is used for the high-performance network (see section VII for details on the network architecture).

Dell Force10 S55. The Dell Force10 S55 switch is a 1/10GbE top-of-rack switch optimized to lower operational costs while increasing scalability and improving manageability at the network edge. Two of the seven Dell Force10 S55 switches are highlighted in Fig. 4.2. Each switch has the following specifications:

- ▷ 44× 1GbE ports
- ▷ 2-port 12Gbps high-speed stacking module
- ▷ 2-port 10GE SFP+ module

Dell Force10 Z9500. The Dell Force10 Z9500 switch is a high-density 3U switch with 132 40GbE ports (528 ports of 1/10GbE using breakout cables). It has low latency, low power and high throughput to ensure line-rate performance. The two Dell Force10 Z9500 switches with the following specifications are highlighted in Fig. 4.2.

- ▷ 132× 40GbE ports
- ▷ VLT protocol

V Power consideration

The power redundancy follows a 2N scheme allowing the infrastructure to afford a failure of up to half the power supply units (PSUs). As a consequence, each PSU must be able to handle the power demand of the node. Based on an accurate estimate of the peak energy consumption, the head- and management-nodes were equipped with two 750W PSUs and the compute-nodes with two 1100W PSUs. Each PSU was connected to a power distribution unit (PDU), which supplies the power to the whole rack. To satisfy the 2N redundancy scheme, each PDU must be able to handle the load of the entire rack that can reach up to *ca.* 20 kW (racks A1 and A5). As depicted by the green and blue lines in Fig. 4.2 the PDUs are supplied with two independent power lines coming from an uninterruptible power supply (UPS) system. The UPS can provide up to *ca.* 20 minutes of power when the load of Arran reaches full capacity.

VI Storage consideration: `/home/` and `/scratch/`

Two types of storage were defined in Arran, referred to as:

- ▷ the `/home/` storage
- ▷ the `/scratch/` storage

I. /home/ storage. The `/home/` hosts the data generated by users. They are stored on the MD3400 storage array. As described in section IVB., the enclosure has an overall capacity of 12× 4TB, providing an effective space of 30TB when configured in RAID6, with an `ext4` file-system. User directories are automatically mounted on an as-needed basis, through `automount`, the program used to configure a mount point for `autofs`. The `/home/` are mounted only as they are accessed, and are unmounted after a period of inactivity.

The `/home/` is mirrored on two different storage enclosures as shown in Fig. 4.2. The replication is handle by DRBD,[236] a distributed replicated storage system for the Linux platform. Working in tandem with Pacemaker,[237] each storage array is considered as a DRBD block device which can be started, stopped, promoted and demoted. In addition to this redundant configuration, the `/home/` storage is fully backed-up once a week on Dalwhinnie. In addition, a daily incremental back-up supplements the full back-up. ASG-Time Navigator[238] software is used for the backup.

Mixing hardware and software redundancy, the overall architecture provides a robust solution for high data availability.

II. /scratch/ storage. The `/scratch/`, located on the compute-nodes, is used to store checkpoints, integrals, wave-functions, etc. required by the calculations at running time. In this definition, this particular storage is temporary: it is created at job submission via a `prolog` script and is deleted 10 days after job termination by a `cron` script automatically executed once an hour.

The `/scratch/` is further divided into local and distributed storage. As suggested by its name, the local `/scratch/` consists of the total amount of available disk space locally, *i.e.*, on the compute-nodes. As mentioned earlier in section IVA., the compute-nodes have 10× 1.2TB of overall capacity. However, because calculations can run over several months, the local `/scratch/` is configured in RAID5 requiring that all drives, but one, be healthy to operate. Consequently, an effective temporary storage of 10TB (`ext4` file-system) can be allocated per node. Since the operating system (OS) is installed on the same RAID layer, two partitions were created: one hosting the OS and one the `/scratch/`, which was mapped via the Linux Logical Volume Manager (LVM).[239]

The GlusterFS scalable network file-system[240] is used for the distributed temporary `/scratch/` storage. A total of 20 nodes share storage capacity over network. Namely 200× 1.2TB HDDs are inter-connected to form a common `/scratch/`. Each of these 20 compute-nodes is configured in the same way than the compute-nodes with a local `/scratch/`: RAID5 layer with two partitions. In the GlusterFS scale-out language, each RAID5 virtual disk is seen as a brick. To optimize the I/O usage, data are striped across bricks in the volume. For the sake of simplicity, four striped volumes are illustrated in Fig. 4.3. In addition, data are spread randomly across the bricks in the latter distributed volume, yielding the distributed striped volume as illustrated by the two green boxes in Fig. 4.3. Since a brick failure in a distributed striped volume can result in a serious loss of data, the files are replicated. As shown in Fig. 4.3, File1 and File2 are simultaneously stored and/or accessed on two distinct volumes, leading to the distributed striped replicated volume. In this configuration, the overall scale-out `/scratch/` storage amounts 97TB per replica.

It is important to emphasize the fact that the configuration detailed here allows each brick to loose a drive thanks to the RAID5 under-layer and each distributed striped volume can loose a brick through the concept of replica. Such a two-level redundancy provides a low failure probability of the distributed `/scratch/` storage.

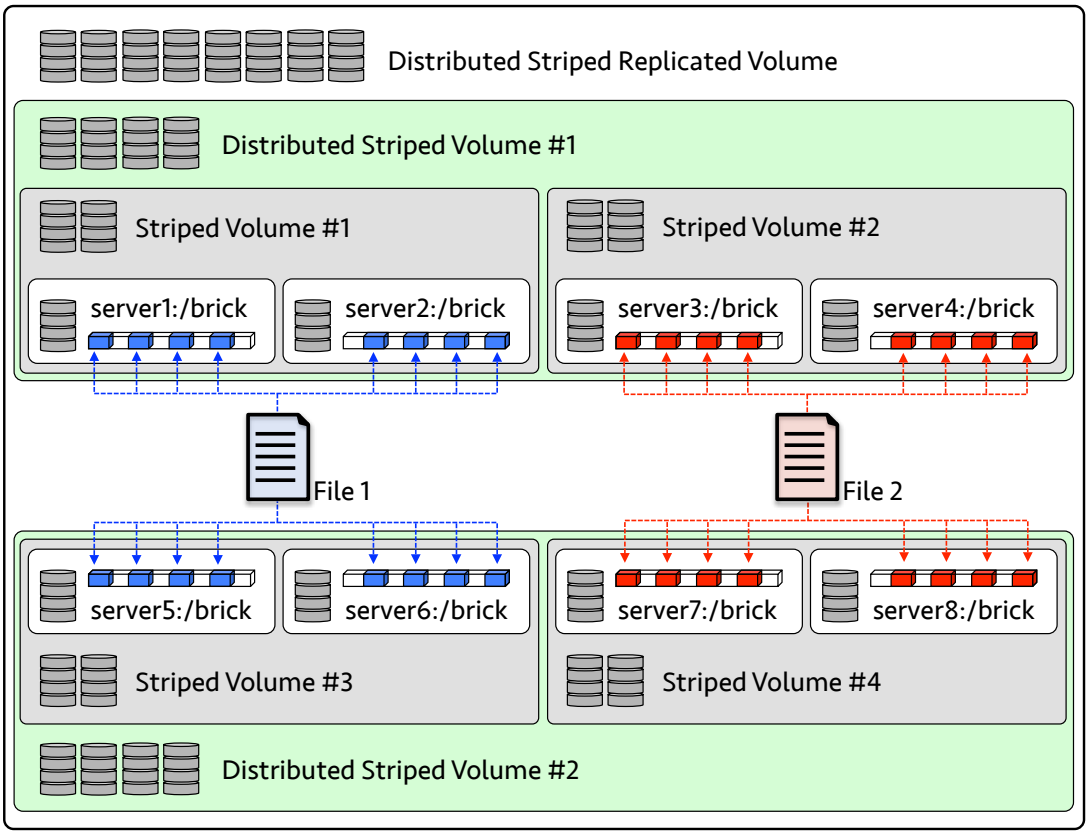


Figure 4.3: Simplified layout of the distributed striped replicated volume used for the scale-out `/scratch/`.

At this point, it is worth mentioning that based on unpublished results and on on-going work (see section XII), the number of available core on the GlusterFS compute-nodes is reduced from 48 to 44. As a matter of fact, the GlusterFS deamon can use up to *ca.* 300% of the CPU. Hence, each GlusterGS compute-node has 4 cores allocated to the deamon, ensuring a smooth running of the distributed storage.

VII Network architecture

The nodes are interconnected following a merged double star topology network, ensuring high level of redundancy. Of the three networks configured on Arran – (i) management network, (ii) iDRAC network, and (iii) high-performance network – a minimum interconnect speed of 10 Gbps is guaranteed on the production network, which is the network used for the calculations.

We will now consider each of the three networks separately.

Management network

The management network is used for the deployment of the configuration, for the OS installation, ugrades and updates and for administration tasks. This network goes through the Dell Force10 S55 switches which are stacked following the so-called daisy chain loop: the cabling starts from the first switch, walks through the other devices and reaches the last unit which is then connected back to the first unit. This configuration ensures a full service continuance in case of a cable and/or a switch

failure. Furthermore, it provides a redundant path to every unit from separate locations, avoiding a connectivity loss to multiple racks. The bottom frame of Fig. 4.4 shows in details the stacking lines in a daisy chain loop.

iDRAC network

The iDRAC is an integrated Dell Remote Access Controller (iDRAC) with an embedded Lifecycle Controller in every server. It provides functionality helping with deployment, firmware updates, monitoring of the hardware and maintaining the nodes. Because it is embedded within each server from the factory, the iDRAC does not need operating system or hypervisor to work, which is of utmost importance for the initial bootstrapping. This iDRAC network, further depicted in dark green in Figs. 4.7, 4.6, and 4.5, is provided by the stack of seven Dell Force10 S55 switches (bottom frame in Fig. 4.4).

In order to accommodate both the management network and the iDRAC on the same stack, each unit of the stack is virtually split into two parts:

- ports 0 – 21 are on the iDRAC network
- port 22 – 43 are on the management network

High-performance network

The high-performance network (HPN), or production network, is used by the resources manager, MPI, YP, NFS mounts, etc.. All the calculations and GlusterFS communicate over this network which goes through the two Dell Force10 Z9500. Redundancy is provided on the data link level, through link-aggregation which combines multiple data links leading to the so-called port trunking, link bundling, bonding, NIC teaming, etc.. In this case, the aggregation is provided by a proprietary protocol known as Virtual Link Trunking (VLT). VLT is available for the enterprise-class network switches and is developed by dell. The protocol creates an aggregation link between the two Z9500 to yield a single virtual entity. In this particular case, the interconnection between the two Z9500 switches is done over four 40 Gbps ports through direct attached cables (DAC), as illustrated in the top frame of Fig. 4.4.

Because of the 10 GbE interfaces on the servers, the network speed of the HPN is limited to 20 Gbps.

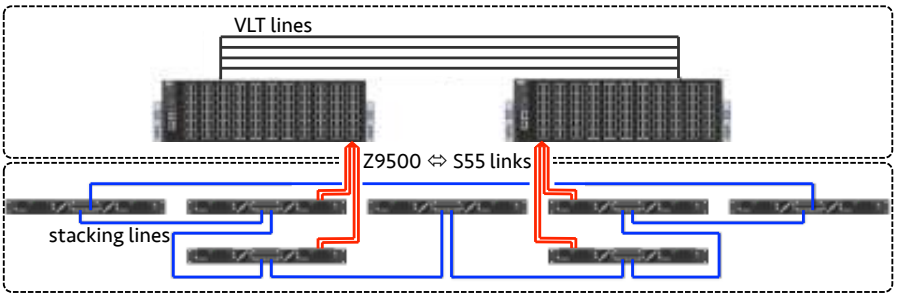


Figure 4.4: The top frame (2× Dell Force10 Z9500 switch) associated to the high-performance network (HPN). In black the 4× 40Gbps VLT-lines. The bottom frame shows the daisy-chain loop stacking over 7× Dell Force10 S55 switches associated to both the iDRAC and the management network. The inter-switch communications is highlighted in red.

As shown in Fig. 4.4, the stack of seven Dell Force10 S55 and the HPN are interconnected through split direct-attached cables (the 2× four red lines). For the sake of redundancy, each Dell Force10 Z9500 core switch is connected to two stack units, such as:

▷ Z9500 rack-a2 → stack units 1 and 2

▷ Z9500 rack-a4 → stack units 4 and 5

Node connectivity

Since the three types of nodes (head-nodes, management-nodes and compute-nodes) run different services, they each have specific connectivity to the networks.

I. Head-nodes. The connectivity of the head-nodes arran1 and arran2, depicted in Fig. 4.5, is the most complex amongst all. First, each node is connected to the iDRAC network and to the management network (green and blue lines, respectively). Second, since the head-nodes host all the production services, *i.e.* resource manager, MPI, executable, etc., they are linked to the HPN, as highlighted by the red and orange lines (see Fig. 4.5). To increase reliability of connectivity, each node is attached to each DELL FORCE10 Z9500 switch, via a separate PCI card, such as:

▷ PCI-I of arran{1,2} → Z9500 rack-a2

▷ PCI-II of arran{1,2} → Z9500 rack-a4

Last but not least, both nodes are linked to the TJU uplink with a public IP to each interface. A third IP, which is resolved by the DNS is attributed to Arran. This IP, a virtual IP (VIP), floats between arran1 and arran2. In our case, the VIP is open to the outside world and is used as log-in point. To minimize the risk of loosing connection to the *virtual* Arran, the two head-nodes are directly connected, This connection is illustrated by the two "arran1 ⇔ arran2" black lines in Fig. 4.5. For an increase redundancy, the two links are attached to the two different PCI cards. Moreover this interconnection is used by DRBD and Pacemarker. In case of a server failure, resources are migrated from the unhealthy node to the sane node. Such configuration ensures a high availability of the head-node along with their associated services.

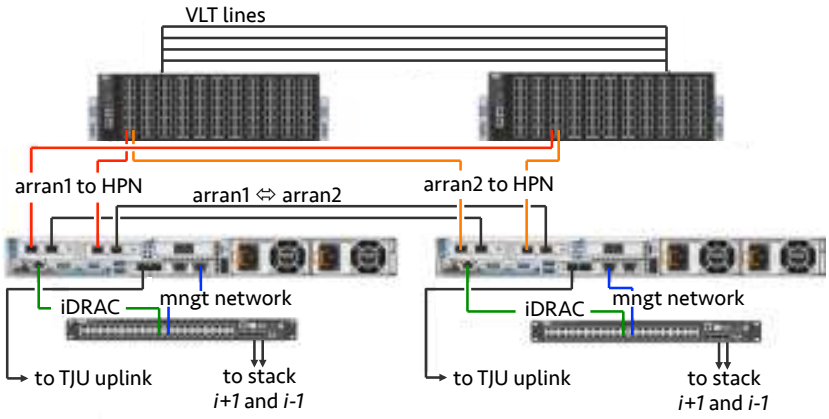


Figure 4.5: Network diagram of the head-nodes, arran1 and arran2. The management network is highlighted in blue, the iDRAC in green and the HPN in red and orange. The inter-head-nodes connections, the uplinks, the VLT lines and the stacking lines are shown in black.

II. Management-nodes. Fig. 4.6 shows the connectivity of the management-nodes. Each node is connected to the iDRAC network and to the management network (green and blue lines, respectively). In addition, the management-nodes have an interface on the iDRAC network in order to monitor/update all other servers. This is referred to as the service processor network (SPN) depicted in light green.

Both arranmngt1 and arranmngt2 are linked to the TJU uplink and a public IP is configured on each interface. A third IP was attributed to the management-nodes: a VIP. As was the case for the head-node, the VIP is resolved by the DNS and is open to the outside world.

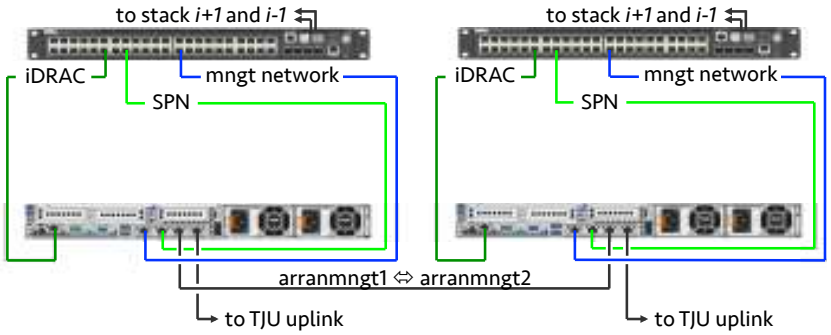


Figure 4.6: Network diagram of the management-nodes, arranmngt1 and arranmngt2. The management network is highlighted in blue, the iDRAC in green and the SPN in light green. The inter-management-nodes connection, the uplinks and the stacking lines are shown in black.

III. Compute-nodes. The compute-nodes – hpn-compute-aX-Y (X being the rack number and Y the mounting unit in rack aX) – are on the iDRAC network and on the management network as shown in Fig. 4.7. The two 10GbE interfaces are on the HPN, enabling the possibility to create link aggregations over the VLT channel. Consequently, in case of a switch failure an inter-compute-node connectivity speed of 10Gbps is guaranteed. The speed under normal condition reaches 20Gbps.

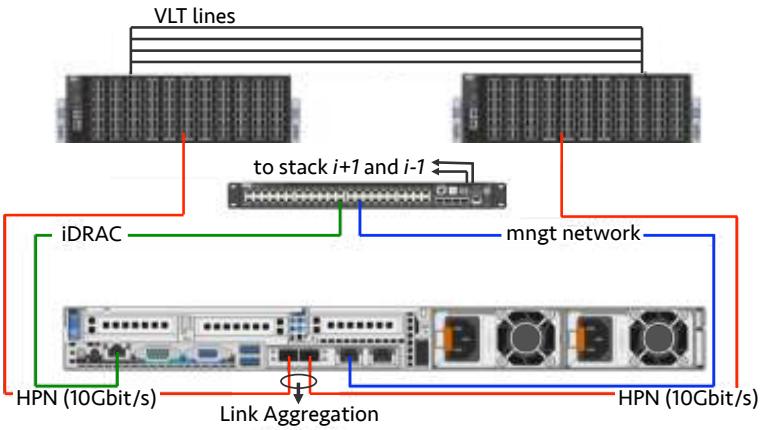


Figure 4.7: Network diagram for the compute-nodes, hpn-compute-aX-Y (X being the rack number and Y the mounting unit in rack aX). The management network is highlighted in blue, the iDRAC in green and the HPN in red. The VLT lines and the stacking lines are shown in black.

VIII Resources management

The resources manager system is a critical component needed to harness such an infrastructure. It performs crucial tasks such as (i) temporally allocate exclusive and/or non-exclusive access to users, (ii) provide a framework to start, execute, and monitor work on the set of allocated resources, and (iii) arbitrate conflicting requests for resources by managing queues and pending work. Among the currently available resources manager (also referred to as work-load manager), all the allocatable resources of Arran – cores, RAM, `/scratch/` storage, etc. – are handled by the open-source manager Simple

Linux Utility for Resource Management (SLURM).[241] SLURM consists of a local deamon running on each compute-node and of a central deamon running on the head-node, Arran. The former reads the common SLURM configuration files and waits for work, executes works, returns status, and waits for more work. The central deamon monitors the state of all compute-nodes and ensures that the compute-nodes have the prescribed configuration before being considered available for use. The central deamon has additional tasks to manage the partitions and queues in accordance to the rules set up for each account.

I. Partitions. The partitions are used to virtually split Arran into separate parts to meet the individual requirement of each user-group. Three partitions were created with SLURM:

- ▷ the *normal* partition counts 4608 cores, with a local `/scratch/` storage of 10 TB per node.
- ▷ the *hfiles* partition counts 880 cores, with a scale-out `/scratch/` storage of 97 TB over 20 nodes.
- ▷ the *huawei* partition counts 48 cores and is used for compilation, linkage and short testing. It has a limit of 90 minutes per session, for all the users.

II. Accounts. Five accounts were created, where each of them enforces restrictions of resources on a set of users. They are typically specified at job submission. `sacctmgr` is used to manage, view and modify the five accounts including the actual 24 users. The account information is stored in a database with the interface being provided by `slurmdbd` (Slurm Database daemon).

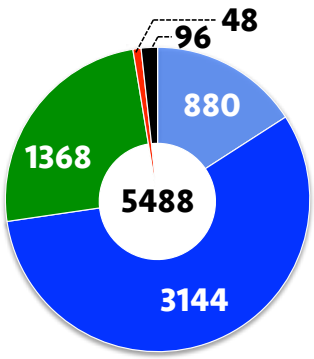


Figure 4.8: The five accounts managed by `sacctmgr`.

The first account (Acc-I, in green in Fig. 4.8) is shared between two research groups and consists of 1368 cores and 18 users. This account uses the *normal* queue as default.

Both Acc-II and Acc-III (in blue and in light blue in Fig. 4.8, respectively) are owned by a single research group. Acc-II counts 3144 cores for five users with the *normal* queue as the default. Acc-III counts 880 cores for five users and the *hfiles* queue sets as default. Since one of the users runs large parallel job requesting a common `/scratch/` space, only the *hfiles* partition can be used. Therefore, this particular user has illimited time with higher priority than the four other users, who are restricted to 6 hours per job.

Acc-IV (in red in Fig. 4.8) is used for compilation, linkage, and short testing. It consists of 48 cores, has a time limit of 90 minutes and all 24 users can use it. The *huwaei* queue is set as default. This account is in place to ensure a continuously available set of resources. Indeed, compilation and interactive work

is strictly forbidden on the head-node.

Acc-V (in black in Fig. 4.8) consists of 96 cores. This account with the *normal* queue as default is meant to be used by visiting Professors, groups, or for scientific collaborations.

III. SLURM configuration. Since arran1 and arran2 works in a *master-slave* configuration, the backup controller is unused, yet specified. Indeed, in a failure event of arran1, all the resources (database keeping track of the work status, `/home/`, etc.) are moved to arran2 via Pacemaker.

For brevity, the key-points of the SLURM configuration files are the following:

- ▷ default memory of 1024 MB per allocated core
- ▷ maximum of 48 task per node, which corresponds to a single task per thread
- ▷ `SelectTypeParameters` is set to `CR_Core_Memory` making cores and memory consumable resources

Note that on hyper-threaded systems, SLURM counts each thread as a CPU to satisfy a job resource requirement.

Association-based enforcement was set up in a way that no new job is allowed to run unless a corresponding association exists in the system. In addition, users are limited by their respective association limits, defined by the accounting manager.

`cgroup` is the plugin for process tracking on a job step basis. This plugin mechanism identifies all children processes and spawns for a user job step. The kernel `cgroup` patch provides the minimum essential kernel mechanisms required to efficiently track processes. It has a minimal impact on the system and provides hooks for specific subsystems or any additional desired behavior.[242] Every 30 seconds, the `cgroup` plugin collects information on the maximum RAM usage, the swappiness, the soft and hard memory limits, the kernel memory usage, and the CPU usage to ensure each job does not exceed the limit enforced at submission time.

IX Mass deployment

For an infrastructure consisting of 116 compute-nodes and a planned extension to 212 compute-nodes, automated tools for "unattended" mass deployment of Linux, installation, configuration and customization, are key players: FAI, Fully Automated Installation,[243] and Ansible[244] were used for the deployment and configurations of Arran.

The installation boostrapping process starts with running a `racadam` script. The purpose of this script is to customize the PowerEdge R630 to meet our needs. It includes configuring the iDRAC interface, collecting the MAC addresses, changing the iDRAC username and password, creating DHCP and host files entries, temporary setting up PXE boot to start FAI, disabling the hotspare to balance the power load on the two PSUs, and setting the thermal profile to maximum performance, the offset of the fans to +25%, a minimum fan speed of 60%, and the maximum exhaust temperature to 50°.

A. FAI

FAI is a non-interactive system to install, customize and manage Linux systems and software configurations on computers. It is a tool for unattended mass deployment of Linux. Starting from a virgin

server, the systems were installed and completely configured to meet the exact needs, without any interaction necessary.

The compute-nodes are installed over the management network from the configuration files located on the management-node. Based on the hostname attributed at boostrapping, the process starts with setting up the class of the compute-node being installed. Once the class is defined, FAI proceeds to the installation and starts with partitioning the virtual RAID5 disk via `parted` for either local or distributed `/scratch/` storage. Then, FAI installs, updates and upgrades the operating system, Ubuntu 14.04 LTS in this case. The next step involve the installation of 152 packages required for the configuration of the cluster. Finally, the GRUB boot loader package is installed and configured. The latter package is essential to select a specific kernel configuration at boot time. Indeed, in case of incompatibility of a functional requirement arising from a kernel update, one would desire to downgrade without re-installing.

Due to the large number of packages to download and because the Chinese internet speed is rather low and unreliable, a proxy server is configured on the management-node.

B. Ansible

Ansible is an agentless end-to-end IT application which sets up the hosts through the use of playbooks, plain texts written in Ansible Automation language. The playbooks describe the desired end-state and contain plays, which contain tasks called modules. The latters run on a set of pre-defined hosts listed in an inventory.

Ansible is used to orchestrate the infrastructure and manage the network and OS. In particular, Ansible runs sequentially a set of roles, each of them consisting of several playbooks. In the case of Arran, we defined seven roles installing, deploying and customizing a large set of different tools needed to meet the requirement of our end-state architecture.

- ▷ the `common` role installs `ntp` on the nodes, deploys the host file in `/etc/hosts` and configures the swappiness, a Linux kernel parameter which controls the relative weight given to swapping out runtime memory. The `common` role also has specific playbooks for the compute-nodes, including deployment of the `cron` job to clean up the temporary `/scratch/` storage, install the Linux Environment Module package which provides a dynamic modification of a individual environment. Finally, the permissions of the `/scratch/` directory are changed to allow users to read/write and execute.

- ▷ the `slurm` role configures the work-load manager SLURM. It starts with the installation of all the required packages though `apt-get install`, copies keys, creates and changes the directory permissions used for logging to keep track of all the jobs, and deploys the SLURM configuration files and finally creates all needed directories and symlinks.

- ▷ the `network` role creates the link aggregation on the compute- and on the head-nodes. The configuration consists of 18 lines added to `/etc/network/interfaces`.

- ▷ the `nis` role installs ans configures the Network Information System[245] (`NIS`) which is a network naming and administration system for smaller networks developed by Sun Microsystems. Using `NIS`, each host client or server computer in the system has knowledge about the entire system. A user at any host can get access to files or applications on any host in the network with a single user identification and authentication. The configuration sets arran1 as master and arran2 as

slave which only has copies of the `NIS` databases and receives these copies from arran1 whenever changes are made to its databases. The compute-nodes are configured as `NIS` clients.

- ▷ the `nfs` role installs `nfs` server and creates all the directories: `/srv/nfs/apps`, `/srv/nfs/home` and `/srv/nfs/slurm-llnl`.
- ▷ the `motd` role deploys the message of the day which is seen upon log-in on the hosts.
- ▷ the `Check_MK` role enables the monitoring software Check_MK[246] (see X) from the configuration file.

X Health checks

Several tools are in place to check the health status of Arran. The main role of the latter tools is to guarantee that Arran is *healthy* enough to host new jobs without compromising any part of the infrastructure. In this regard, Check_MK and various scripts called by the scheduler SLURM are used to monitor the health status of Arran.

I. Check_MK. Check_MK is a comprehensive IT monitoring solution in the spirit of Nagios. It covers all important areas: from the monitoring of applications, operating systems, hardware, networks and processing centers. In order not to interfere with the HPN, the checks run over the management network and over the iDRAC network.

On the management network, the own Check_MK check-plugins are used to monitor the CPU usage and load, the disk I/O, the mounted partitions (*i.e.* `/boot`, `/scratch/` and `autofs`), the status of the interfaces (*i.e.* link aggregation to the HPN from the compute-nodes, interfaces on the iDRAC and on the management network), the kernel, the logs, the memory usage, etc.

On the iDRAC network, Check_MK reports the status of the hardware spanning amperage, voltage, batteries, cooling, fans, memory DIMM sockets, physical disks, temperature sensors, CPUs, etc.

Overall a few 13'168 checks are performed every minutes on the cluster.

II. Epilog script. Every job submitted through SLURM is associated to an epilog script, which is called at job completion. The epilog script has two roles:

- ▷ clean the eventual leftover processes from the run. Since no running process cleans the SysV IPC semaphores and the SHM segments, it is necessary to run such a script on a regular basis. This is particularly important for GAMESS jobs since it tends to leave many allocated semaphores and SHM segments. Consequently, no new `ddikick.x`* process can be started on the compute-nodes.
- ▷ clean the semaphores. In case of a job failure, it happens that SLURM is not able to clean all the processes. Therefore, the script purges all remaining children processes associated with the parent job. Special care is needed not to kill root processes or system daemon jobs.

III. Health check script. In addition to the epilog script, a health check script periodically runs on all allocated and idle compute-nodes. It checks every minutes the availability of the `/scratch/` storage. Even though Check_MK reports the status of the `/scratch/`, no action is performed in case

*`ddikick.x` is the GAMESS kickoff program used for DDI running over TCP/IP sockets

of a detected failure. The `health_check` tests (i) if the `/scratch/` is correctly mounted and (ii) if the user has the correct read/write permissions. If any of the latter tests fail, SLURM updates the state of the node from the current state to a DRAIN state and automatically reports that the `scratch partition is not mounted or working properly`.

XI Some facts

The common computational chemistry packages available through the Linux `modules` are:

- ▷ GAMESS
- ▷ Columbus
- ▷ LAMMPS
- ▷ Gaussian
- ▷ Quantum Espresso
- ▷ Theodore
- ▷ Turbomole
- ▷ Berkely GW

Visualization tools, such as GaussView and Molden, are accessible through `X11` fowarding. All these packages are compiled with a variety of libraries and compilers wich are also available via the Linux `modules`:

- ▷ Intel compilers (versions 2015 and 2016)
- ▷ GNU compilers (versions 4.6 and 4.8)
- ▷ PGI compilers (version 15.10)
- ▷ Intel MKL (versions 2015 and 2016)
- ▷ hdf5 (version 1.5)
- ▷ Intel MPI (version 5.1)
- ▷ OpenMPI (version 1.8)
- ▷ MPIch (version 15.10)
- ▷ Global Array (GA4.3 and GA5.4)
- ▷ fftw (version 2.1 and 3.3)

Basics instructions on how to write a submission script are given to the users. For GAMESS and Gaussian jobs, the in-house developed `qgms` and `qg09` submission scripts are available via `modules`.

XII Performance study of a scale-out GlusterFS storage

With high throughput as ultimate goal, a performance study of the scale-out GlusterFS `/scratch/` storage is performed. In a first time, the disk usage of different quantum theory codes was established. The read/write ratio as a function of the total job size was obtained via the statistics provided by `nfsstat`, and is summarized in Fig. 4.9.

In a second time, the performance of GlusterFS is evaluated using FIO 2.1.3, an I/O tool meant to be used both for benchmark and stress/hardware verification. The nodes used in the tests run the Linux distribution Ubuntu 14.04, and GlusterFS 3.4.2. The hosts are interconnected via a 20 Gbps Ethernet network. Each node has two Intel Xeon CPU E5-2690 v3 (12 cores, 2.6GHz), 256 GB of LR-DIMM 2133 MT/s DDR4 physical memory, and ten 1.2 TB 10 K-RPM SAS disks.

All the tests follow a mixed random reads and writes I/O pattern issued with the Linux native asynchronous I/O, known as `libaio`. Based on the behavior of the codes depicted in Figure 4.9, seven tests were designed, as summarized in Table 4.1.

	test 1	test 2	test 3	test 4	test 5	test 6	test 7
test size [MB]	1'024	28'672	45'056	67'584	227'328	436'224	544'768
reads [%]	0	0	1	5	65	71	85
writes [%]	100	100	99	95	35	29	15

Table 4.1: Read/write ratio and size of the seven tests.

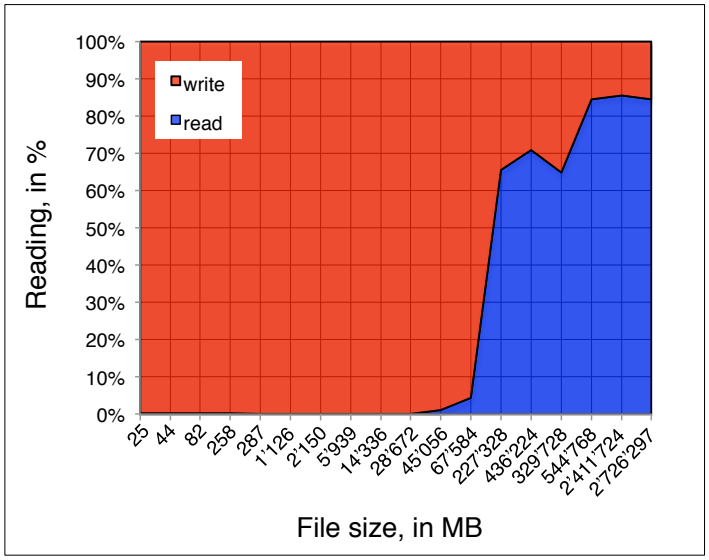


Figure 4.9: Read/write ratio at different file size (in MB). In blue the reads and in red the writes percentage.

Each of the tests described in Table 4.1 was submitted on different nodes configuration to study the performance of the GlusterFS scale-out storage. In this regard, seven nodes were setup with the following specifications:

- ▷ node-1: 1× bare 1.2 TB disk
- ▷ node-2: 1× 12 TB RAID0 virtual disk
- ▷ node-3: 1× 1.2 TB brick
- ▷ node-4: 1× 12 TB RAID0 virtual distributed brick
- ▷ node-5: 10× 1.2 TB distributed bricks
- ▷ node-6 & -7: 2× [1× 12 TB RAID0 virtual] distributed bricks

The speed results collected over two runs are summarized in Table 4.2 and 4.3.

	test 3 r (1%)	test 4 r (5%)	test 5 r (65%)	test 6 r (71%)	test 7 r (85%)
node-1	54	147	638	525	12222
node-2	398	598	888	845	930
node-3	172	1742	2219	1000	978
node-4	409	1892	3949	1233	1349
node-5	176	1826	2275	1019	994
node-6 & -7	482	1861	6665	1082	980

Table 4.2: Read speed of the different scenarios on the seven tests (see 4.1). Results are reported in KB/s.

At file size smaller than *ca.* 200 GB, the read speed is considerably smaller for the bare disk (node-1) than for the virtual RAID0 12 TB disk (node-2). As a matter of fact, a ratio of 4 to 7 is observed in

tests 4 and 7, respectively. For large files (*i.e.* test 5 and beyond) even though the difference decreases node-2 gives the fastest read speed.

For the four nodes configured with GlusterFS (nodes-3 to 7), it appears that the size of the brick is the limiting factor on the read speed. A comparison of the measured speed for nodes 3 and 5 and for nodes 4 and 6 & 7 illustrates this observation: for the five measurements the speed for each set of two nodes is very similar.

As was the case for the non-GlusterFS nodes (nodes 1 and 2) the PERC controller handling the RAID0 virtual disk seems to fasten the read speed. Indeed, nodes-4 and 5 with RAID0 underneath GlusterFS have greater speed than the corresponding non-GlusterFS nodes.

	test 1 w (100%)	test 2 w (100%)	test 3 w (99%)	test 4 w (95%)	test 5 w (35%)	test 6 w (29%)	test 7 w (15%)
node-1	1528627	1499750	5410	2813	343	214	137
node-2	1520230	1500774	39350	11370	478	345	164
node-3	47701	42691	42146	33119	1195	408	172
node-4	44757	46244	40498	35979	2127	504	238
node-5	46227	41773	39106	34724	1225	416	175
node-6 & -7	45526	51815	47710	35383	3590	442	173

Table 4.3: Write speed of the different scenarios on the seven tests (see 4.1). Results are reported in KB/s.

Table 4.3 reports the write speed of the different scenarios investigated. The speed of the non-GlusterFS nodes 1 and 2 shows a drastic drop from tests 1 and 2 to test 3. When comparing tables 4.2 and 4.3 similarities are quite obvious: the virtual RAID0 node-2 displays greater write speed than the bare disk. GlusterFS-nodes outperforms non-GlusterFS nodes already at small file size, *e.g.*, *ca.* 45 GB.

XIII Concluding remarks and further development

The two-level redundancy used throughout the physical design and software implementation provides a low failure probability, maximizing the availability of Arran. Such stability is easily proven by several calculations which exited normally after a few 80 days of computations spread over several nodes across the cluster.

It is to be mentioned that the 13'168 checks made every minute guarantee an efficient way to troubleshoot issues arising at any level: hardware, network, software, etc. The health checks are also of utmost important to automatically update the compute-node status preventing jobs to start on a defective compute-node. In a disaster scenario requiring bare metal (re-)installation the tools used for mass deployment allows (i) an easy-way for re-installation and (ii) future extension, which is critical in a fast-pace growing environment.

The exhaustive on-going study of the scale-out GlusterFS `/scratch/` storage allows a fine tuning of the brick size and number to maximize the read and write speed while maintaining high redundancy. It also reports an across-the-board increased read/write speed in the range of the file sizes generated by common quantum chemical codes.

Amongst the unchecked items in the *Arran-wish-list* one should mention:

- ▷ the planned extension of Arran from 5568- to 10944-hyperthreaded cores.
- ▷ GIT a version control system that is used for software development and other version control tasks. Such tools is greatly appreciated in case of incompatibility between an update and critical components of the cluster.
- ▷ giving the possiblity to users to set up e-mail notification at job completion.
- ▷ e-mail notification from Check_MK in order to increase the response speed of the system Administrator.
- ▷ configure the qualtiy of service (QOS) in SLURM. Although this is partially configured, one needs to monitor the behaviour of the scheduler and fine-tune the configuration file to ensure a fair usage of the resources within the five different accounts.
- ▷ work on a easy way to set up GlusterFS from Ansible. The problem lies in the fact that the tuning GlusterFS is system dependent. In particular many commands are executed only once, at bootstrapping, on a specific set of compute-nodes. In addition, in the above mentioned disaster scenario, the procedure for brick restoration is case specific.

Chapter 5

Concluding remarks and perspectives

The work of this thesis involved the development, implementation, and application of several cost-effective single reference methods for accurate calculations of molecular structure and properties of real systems. Aiming at high accuracy, scalability and efficiency, efforts in development of new quantum chemical methods, improvement and/or enhancements in existing methods, design of high availability and reliable HPC, applications towards real-time experimental investigations that exemplify the new theoretical treatments, have been undertaken in this thesis work. These aims have been undertaken as laid out in the opening chapter of this thesis and can be summarized as follows.

A. Quantum mechanical challenge

Chapter 2 reports the successful implementation of several double-hybrids DFTs. In total, more than 300 DSD-DFTs were added to GAMESS. All can be easily selected and tuned using a set of new keywords. In addition, the design of new DSD-DFTs is facilitated in such a way that no modification of the source code is required, allowing users to optimize the parameters of the functionals directly from the set of new keywords in the `$DFT` group of the input file.

An exhaustive performance study of methodologies on non-covalent interacting systems was carried out. This study revealed the high-level of performance of the DSD-DFT approach, promoting the double-hybrid scheme in general as a promising technique to account for correlation effects.

In this context, the implementation and performance of a new family of double-hybrids, the SCS-DFTs, is presented. Each of the 300 DSD-DFTs now has a corresponding SCS-DFT. Exemplifying the SCS-DFTs, the SCS-PBEPBE functional with only two parameters was shown to perform as well as most of the DSD-DFTs, which have five parameters. Most importantly, the SCS-DF scheme does not over-weight the correlation energy component to reach high accuracy. To date, only the parameters of the SCS-PBEPBE were optimized on the reduced data set suggested in this work, and its performance was studied on two of the seven new data sets also suggested herein. Further work in developing this new family of double-hybrid would involve (i) an extensive study of this new method on larger chemical systems governed by weak interactions, and (ii) optimization of the parameters for SCS-BPBE, SCS-BP86 and SCS-PBEPW91 as their DSD-version have shown similar performance to DSD-PBEPBE.

On the way towards the implementation of a different approach for including the correlation component of the total energy in the double-hybrid DFT, the attenuated MP2 developed by M. Head-Gordon *et al.* was implemented in GAMESS, such that the users can select between three types of MP2 types (Fig. 5.1).

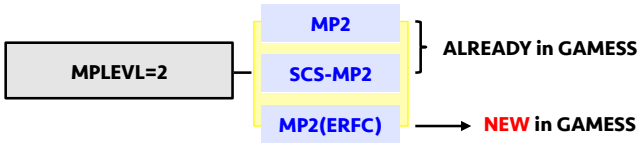


Figure 5.1: New implementation scheme of the MP2 quantum chemical method.

Combining MP2($erfc$) with the approximate GGA correlation funtional leads to a new type of double-hybrids, the DH($erfc$)-DFs. Their successful implementation in GAMESS is reported, and further work would involve the optimization of the exchange-correlation parameters together with the cc-pVDZ and at the cc-pVTZ basis set levels. The rationale for cc-pVDZ and -pVTZ is in the tabulated values for the attenuation parameter ω.[15;156] The procedure used to extrapolate the parameters of SCS-PBEPBE at the CBS limit would be a judicious way to obtain the CBS ω as well as the CBS exchange-correlation

parameters.

Finally, for the sake of cost-effectiveness and to broaden the double-hybrids DFTs range of application, the source code was modified to allow all families of DH-DFTs to run in their corresponding RI-approximation. With a speed up over 50 times, the double-hybrids DFT in their RI-versions show high accuracy at much lower cost. The new RI double hybrids are not only important in terms of saving computational time but also in terms of system requirements, in particular, disk space and memory requirements are significantly reduced. To enable the RI-version of DH($erfc$)-DFs the attenuated three- and two-center integrals would need to be worked out,

$$\langle ij|kl\rangle \approx \sum_{\mu}\sum_{\nu} \langle ij|\mu\rangle \langle \mu|\nu\rangle^{-1} \langle \nu|kl\rangle \tag{5.1}$$

with the Coulomb attenuated operator replacing the standard $^1/_r$ operator.

$$\frac{1}{r} \approx \frac{erfc\,(\omega \times r)}{r} \tag{5.2}$$

B. Hardware challenge

Chapter 4 reports the design, development, and mass deployment of a highly available HPC infrastructure, Arran, for large quantum chemistry calculations. The two-level redundancy used throughout the physical design and software implementation provides a low failure probability, maximizing the availability of Arran. Such stability is illustrated, for example, by several calculations exiting normally after over 80 days of computations spread over several nodes across the cluster.

The designed bootstrapping capabilities and tools used for mass deployment have shown their efficiency and reliability , *e.g.*, several times during the two-week maintenance, which took place in May 2016. Nonetheless, efforts in setting-up e-mail notifications, GIT, QOS in SLURM and a GlusterFS role in Ansible should be made. Notably, the System Administrator would benefit from Check_MK e-mail notifications, as it would allow a shorter response time to start the troubleshooting process. GIT would allow full control over the system versioning: in case of incompatibility between an update and critical components of the cluster, the system could be downgraded to a previous version and, therefore, such tools would be greatly appreciated. Configuring GlusterFS through Ansible would require extension the exhaustive on-going study of the scale-out GlusterFS `/scratch/` storage to additional scenarios, among which should be mentioned (i) the impact of the RAID redundancy on the read/write (rw) speed, (ii) the impact of the file system of the RAID virtual disk on the rw speed, (iii) the impact of striping and replicating on the rw speed are primary focus. Furthermore, the scalability should be studied in order to eventually extend such partitioning to the whole cluster.

Finally, the overall performance (in term of Flops) of the cluster should be assessed by solving numerical linear algebra through the software library LINPACK. However, based on the CPU frequency, the number of cores per CPU, and the number of flop per cycle, 580 TFlops would be a good estimate of the theoretical peak performance. Assuming an 80% cluster effectiveness, Arran would rank 250–300 in the TOP500 list.

C. Chemical challenge

Chapter 3 illustrates a successful synergy between theoretical predictions and experimental observations. The investigation discussed relates the competition between pentaindenocorannulene (pIC)

self-assembly and its aggregation with C_{60}. The work brings together the efforts in algorithm design and high performance computing enablements, with that of experimental procedures, to offer a higher level of predictability than would be otherwise enabled.

Both theory and experiment reveal (i) a new stacking pattern of pIC, following a columnar motif, (ii) cyclic voltammetry with reversible multi-electron reduction profile (iii) NMR spectra dependence on the concentration suggesting that pIC is prone to form dimers and higher self-assemblies in solution, and, (iv) a preference for the "nest" C_{60}@pIC_2 arrangement. Furthermore, neither experimental nor theoretical UV-vis data showed any new absorption bands for the aggregate.

As computing power increases, more accurate methods become affordable and/or larger system can be studied, which in turn leads to an increased number of synergistic investigations. In particular, insights gained from both experiment and theory theory fill important gaps in our knowledge and facilitates more detailed analysis of raw results. A good example is the structure analysis of the aggregate C_{60}@pIC_2. Thanks to experimental evidence obtained in solution from the method of continuous variations applied to the ^{1}H NMR chemical shifts of different molar fractions of C_{60} and pIC at a constant total concentration of 2 mM, a 2:1 stoichiometry was suggested. This restricted the structural search to only four starting geometries. The relative stability of the relaxed complexes allowed assignment of the exact orientation of C_{60} within the heavily disordered crystal, as depicted in Fig. 5.2.

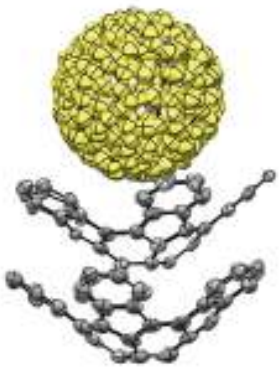


Figure 5.2: Crystal structure of the aggregate C_{60}@PIC_2.

Last but not least, the level of theory applied to this investigation was enabled by Arran, the high performance cluster that we built, and such level is significantly higher than most infrastructure would have enabled.

Chapter 6

Acknowledgments

First, I would like to thank Prof. Dr. Kim K. Baldridge for supervising my Ph.D. thesis. I am very grateful for her trust, support and help over the last years. She introduced me to code development and scripting. She also gave me the great opportunity to design, build and manage a large-scale HPC center. The same is true for her husband, Prof. Dr. Jay S. Siegel.

I thank my committee members for following the progress of this Ph.D. thesis on a yearly-based presentation, report, and exam.

I thank Tyanko for being so patient in explaining me the principles of a supercomputer from the very basics and for the insightful discussion at La Bamba. We had a great time in Tianjin with the build of Oban, Dalwhinnie, Dalmore and Arran. Naming the clusters after whiskys was such a good excuse to have sip (or two) in a moment of intense stress and/or frustration... *TIC*.

I would like to thank the group members creating a good and friendly atmosphere. I also thank the members of the Robinson group for the nice chats and relaxing moments around a coffee, a birthday cake, a pizza, a beer, ... A special thank to Mülisse and Matu for never being late (!) at the *ca.* 2160 coffee breaks we had together (Mülisse was right: we almost spent a month salary for horrible coffee!). They were always ready for a run, a barbecue, a few drinks at the DoBar, a whisky-cigar night, or for a few Glühwein at the Christmas market.

I wish to thank the graduate school for organizing Aperos, retreats, and Doktorandentag. This created a good atmosphere between the members of the research groups within the CMSZH.

I am thankful to all the friends in Zürich, and in Tianjin who participated in making my stay in both cities pleasant and enjoyable. A particular thank to Torben for his friendship, and for his help in translating the abstract into German.

I thank Helena and my family for their years of support and encouragement.

Bibliography

[1] R. Breslow, M. V. Tirrell, J. K. Barton, M. A. Barteau, C. R. Bertozzi, R. A. B. P. Gast, I. E. Grossmann, J. M. Meyer, R. W. Murray, P. J. Reider, W. R. Roush, M. L. Shuler, J. J. Siirola, G. M. Whitesides, P. G. Wolynes, and R. N. Zare, *Beyond the Molecular Frontier: Challenges for Chemistry and Chemical Engineering.* (National Academies Press (US), 2001).

[2] C. R. Bertozzi, C. J. Chang, B. G. Davis, M. O. de la Cruz, D. A. Tirrell, and D. Zhao, ACS Cent. Sci. **2**, 1 (2016).

[3] I. Shavitt and R. J. Bartlett, *Many-Body Methods in Chemistry and Physics* (Cambridge university press, 2009).

[4] C. Møller and M. S. Plesset, Phys. Rev. **46**, 618 (1934).

[5] J. Čížek, J. Chem. Phys. **45**, 4256 (1966).

[6] J. Čížek, Adv. Chem. Phys. **14**, 35 (1969).

[7] J. Čížek and J. Paldus, Int. J. Quantum Chem. **5**, 359 (1971).

[8] P. Hohenberg and W. Kohn, Phys. Rev. B **136**, 864 (1964).

[9] W. Kohn and L. J. Sham, Phys. Rev. A **140**, 1133 (1965).

[10] S. Grimme, J. Chem. Phys. **118**, 9095 (2003).

[11] A. Szabados, J. Chem. Phys. **125**, 214105 (2006).

[12] S. Kozuch, D. Gruzman, and J. M. L. Martin, J. Phys. Chem. C **114**, 20801 (2010).

[13] S. Kozuch and J. M. L. Martin, Phys. Chem. Chem. Phys. **13**, 20104 (2011).

[14] M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. Su, T. L. Windus, M. Dupuis, and J. A. Montgomery, J. Comput. Chem. **14**, 1347 (1993).

[15] M. Goldey and M. Head-Gordon, J. Phys. Chem. Lett. **3**, 3592 (2012).

[16] J. L. Whitten, J. Chem. Phys. **58**, 4496 (1973).

[17] G. E. Moore, Electronics **8**, 38 (1965).

[18] `http://www.intel.com/content/www/us/en/silicon-innovations/moores-law-technology.html` (2016).

[19] P. G. anb P. Gargiani, G. Parker, and A. Yu, IEEE Spectrum (1989).

[20] C. Engelmann, S. L. Scott, C. Leangsuksun, and X. He, Journal of computers **1**, 43 (2006).

[21] E. Lusk and T. Sterling, *Beowulf Cluster Computing with Linux*, second edition ed., edited by W. Gropp (The MIT Press Cambridge, Massachusetts London, England, 2003).

[22] M. Feldman, "Petaflop club closes in on 100 members," .

[23] E. Yilmaz and L. Gilly, "Redundancy and reliability for an hpc data centre," Partnership for Advanced Computing in Europe (2012).

[24] K. Kharbas, D. Fiala, F. Mueller, and K. F. ans Christian Engelmann, in *Distributed Computing Systems* (2012) pp. 615–626.

[25] A. Szabo and N. S. Ostlund, *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory* (McGraw-Hill, 1989).

[26] F. Jensen, *Introduction to Computational Chemistry*, second edition ed. (Wiley, 2007).

[27] C. J. Cramer, *Essentials of Computational Chemistry, Theories and Models*, second edition ed. (Wiley, 2006).

[28] P.-A. Malmqvist, J. Olsen, P. R. Taylor, J. Almlöf, and R. Ahlrichs, *European summerschool in quantum chemistry, Book I*, edited by P.-O. Widmark (Lund University, 2009).

[29] B. O. Roos, P. R. Taylor, T. Helgaker, and D. J. Tozer, *European summerschool in quantum chemistry, Book II*, edited by P.-O. Widmark (Lund University, 2009).

[30] M. Cossi, V. Barone, and P. R. Taylor, *European summerschool in quantum chemistry, Book III*, edited by P.-O. Widmark (Lund University, 2009).

[31] J. Klimeš and A. Michaelides, J. Chem. Phys. **137**, 1209011 (2012).

[32] C. C. J. Roothaan, Rev. Mod. Phys. **32**, 179 (1960).

[33] A. Szabo and N. S. Ostlund, J. Chem. Phys. **67**, 4351 (1977).

[34] M. O. Sinnokrot and C. D. Sherrill, J. Phys. Chem. A **108**, 10200 (2004).

[35] P. Jurečka, J. Černý, and P. Hobza, Phys. Chem. Chem. Phys. **8**, 1985 (2006).

[36] G. Frenking, I. Antes, M. Böhme, S. Dapprich, A. W. Ehlers, V. Jonas, A. Neuhaus, M. Otto, R. Stegmann, A. Veldkamp, and S. F. Vyboishchikov, *Pseudopotential Calculations of Transition Metal Compounds: Scope and Limitations*, Vol. 8 (VCH Publishers, Inc., 1996).

[37] Y. Jung, R. C. Lochan, A. D. Dutoi, and M. Head-Gordon, J. Chem. Phys. **121**, 9793 (2004).

[38] R. C. Lochan, Y. Shao, and M. Head-Gordon, J. Chem. Theory Comput. **3**, 988 (2007).

[39] A. Tkatchenko, R. A. DiStasio, M. Head-Gordon, and M. Scheffler, J. Chem. Phys. **131**, 094106 (2009).

[40] A. Tkatchenko and M. Scheffler, Phys. Rev. Lett. **102**, 073005 (2009).

[41] A. Hesselmann, J. Chem. Phys. **128**, 144112 (2008).

[42] A. Pitoňák and A. Hesselmann, J. Chem. Theory Comput. **6**, 168 (2010).

[43] E. Runge and E. K. U. Gross, Phys. Rev. Lett. **52**, 997 (1984).

[44] E. K. U. Gross and W. Kohn, Phys. Rev. Lett. **55**, 2850 (1985).

[45] E. K. U. Gross and W. Kohn, Adv. Quantum Chem. **21**, 255 (1990).

[46] M. Feyereisen, G. Fitzgerald, and A. Komornicki, Chem. Phys. Lett. **208**, 359 (1993).

[47] F. Weigend, M. Häser, H. Patzelt, and R. Ahlrichs, Chem. Phys. Lett. **294**, 143 (1998).

[48] F. Weigend and M. Häser, Theoret. Chem. Acc. **97**, 331 (1997).

[49] A. Hellweg, C. Hättig, S. Höfener, and W. Klopper, Theoret. Chem. Acc. **117**, 587 (2007).

[50] R. A. DiStasio, Y. S. Jung, and M. Head-Gordon, J. Chem. Theory Comput. **1**, 862 (2005).

[51] P. Pulay and S. Saebø, Theor. Chim. Acta **69**, 357 (1986).

[52] J. Almlöf, Chem. Phys. Lett. **181**, 319 (1991).

[53] M. Häser and J. Almlöf, J. Chem. Phys. **96**, 489 (1992).

[54] P. R. Taylor, G. B. Bacskay, N. S. Hush, and A. C. Hurley, Chem. Phys. Lett. **41**, 444 (1976).

[55] R. J. Bartlett and G. D. Purvis, Int. J. Quantum Chem. **14**, 561 (1978).

[56] G. D. Purvis and R. J. Bartlett, J. Chem. Phys. **76**, 1910 (1982).

[57] J. Noga and R. J. Bartlett, J. Chem. Phys. **86**, 7041 (1987).

[58] Y. S. Lee, S. A. Kucharski, and R. J. Bartlett, J. Chem. Phys. **84**, 5906 (1984).

[59] M. Urban, J. Noga, S. J. Cole, and R. J. Bartlett, J. Chem. Phys. **83**, 4041 (1985).

[60] K. Raghavachari, G. W. Trucks, J. A. Pople, and M. Head-Gordon, Chem. Phys. Lett. **157**, 479 (1989).

[61] W. Klopper and W. Kutzelnigg, J. Chem. Phys. **96**, 2020 (1991).

[62] P.-O. Löwdin, Int. J. Quantum Chem. **S19**, 19 (1986).

[63] J. P. Perdew, A. Ruzsinsky, J. Tao, V. N. Staroverov, G. E. Scuseria, and G. Csonka, J. Chem. Phys. **123**, 62201 (2005).

[64] D. M. Ceperley and B. J. Alder, Phys. Rev. Lett. **45**, 566 (1980).

[65] J. P. Perdew, Phys. Rev. B **33**, 8822 (1986).

[66] A. D. Becke, Phys. Rev. A **38**, 3098 (1988).

[67] C. Lee, W. Yang, and R. G. Parr, Phys. Rev. B **37**, 785 (1988).

[68] N. C. Handy and A. J. Cohen, Mol. Phys. **99**, 403 (2001).

[69] F. A. Hamprecht, A. J. Cohen, D. J. Tozer, and N. C. Handy, J. Chem. Phys. **109**, 6264 (1998).

[70] J. P. Perdew and Y. Wang, Phys. Rev. B **45**, 13244 (1992).

[71] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[72] D. C. Langreth and J. P. Perdew, Solid State Commun. **17**, 1425 (1975).

[73] O. Gunnarsson and B. I. Lundqvist, Phys. Rev. B **13**, 4274 (1976).

[74] A. D. Becke, J. Chem. Phys. **98**, 5648 (1993).

[75] T. W. Keal and D. J. Tozer, J. Chem. Phys. **123**, 121103 (2005).

[76] C. Adamo and V.Barone, J. Chem. Phys. **110**, 6158 (1999).

[77] T. Yanai, D. P. Tew, and N. C. Handy, Chem. Phys. Lett. **393**, 51 (2004).

[78] Q. Wu and W. Yang, J. Chem. Phys. **116**, 515 (2002).

[79] S.Grimme, J.Chem.Phys. **124**, 034108/1 (2006).

[80] A. Karton, A. Tarnopolsky, J.-F. Lamère, G. C. Schatz, and J. M. L. Martin, J. Phy. Chem. A **112**, 12868 (2008).

[81] J.-D. Chai and M. Head-Gordon, J. Chem. Phys. **131**, 174105 (2009).

[82] J. Zheng, Y. Zhao, and D. G. Truhlar, J. Chem. Theor. Comput. **3**, 569âĹŠ582 (2007).

[83] L. Goerigk and S. Grimme, Phys. Chem. Chem. Phys. **13**, 6670 (2011).

[84] E. Clementi, J.-M. André, and J. A. McCammon, eds., *Theory and applications in computational chemistry: the first decade of the second millennium* (AIP, CONFERENCE PROCEEDINGS,, 2012).

[85] T. B. Adler, H.-J. Werner, and F. R. Manby, J. Chem. Phys. **130**, 130 (2009).

[86] T. B. Adler and H.-J. Werner, J. Chem. Phys. , 130 (2009).

[87] J. A. Pople and W. J. Hehre, J. Comput. Chem. **27**, 161 (1978).

[88] H. F. King and M. Dupuis, J. Comput. Phys. **21**, 144 (1976).

[89] M. Dupuis, J. Rys, and H. F. King, J. Chem. Phys. , 111 (1976).

[90] J. Rys, M. Dupuis, and H. F. King, J. Comput. Chem. , 154 (1983).

[91] X. Ren, P. Rinke, V. Blum, J. Wieferink, A. Tkatchenko, A. Sanfilippo, K. Reuter, and M. Scheffler, New J. Phys. **14** (2012).

[92] B. I. Dunlap, J. W. D. Connolly, and J. R. Sabin, J. Chem. Phys. **71**, 3396 (1979).

[93] J. W. Mintmire, J. R. Sabin, and S. B. Trickey, Phys. Rev. B **26**, 1743 (1982).

[94] O. Vahtras, J. Almlöf, and M. Feyereisen, Chem. Phys. Lett. **213**, 514 (1993).

[95] M. Feyereisen, G. Fitzgerald, and A. Komornicki, Chem. Phys. Lett. **208**, 359 (1993).

[96] F. Weigend, M. Häser, H. Patzelt, and R. Ahlrichs, Chem. Phys. Lett. **294**, 143 (1998).

[97] F. Weigend, Phys. Chem. Chem. Phys. **4**, 4285 (2002).

[98] C. A. Schalley, *Analytical Methods in Supramolecular Chemistry* (WILEY-VCH, 2012).

[99] P. Jurečka, J. Šponer, J. Černý, and P. Hobza, Phys. Chem. Chem. Phys. **8**, 1985 (2006).

[100] J. Řezáč and P. Hobza, J. Chem. Theory Comput. **9**, 2151 (2013).

[101] J. Řezáč, K. E. Riley, and P. Hobza, J. Chem. Theory Comput. **8**, 4285 (2012).

[102] Y. Zhao and D. G. Truhlar, J. Phys. Chem. A **109**, 5656 (2005).

[103] Y. Zhao and D. G. Truhlar, J. Chem. Theory Comput. **1**, 415 (2005).

[104] F. Weigend and R. Ahlrichs, Phys. Chem. Chem. Phys. **7**, 3297 (2005).

[105] S. Tsuzuki, K. Honda, T. Uchimaru, and M. Mikami, J. Chem. Phys. **124**, 114304 (2006).

[106] S. Grimme, J. Comput. Chem. **27**, 1787 (2006).

[107] N. L. Allinger, J. T. Fermann, W. D. Allen, and H. F. S. III, J. Chem. Phys. **12**, 5143 (1997).

[108] R. M. Balabin, J. Phys. Chem. A **113**, 1012 (2009).

[109] J. F. Ogilvie and F. Y. H. Wang, J. Mol. Struct. **273**, 277 (1992).

[110] J. F. Ogilvie and F. Y. H. Wang, J. Mol. Struct. **291**, 313 (1993).

[111] R. Peverati, M. Macrina, and K. K. Balridge, J. Chem. Theory Comput. **6**, 1951 (2010).

[112] T. H. J. Dunning, J. Chem. Phys. **90**, 1007 (1989).

[113] R. A. Kendall, T. H. J. Dunning, and R. J. Harrison, J. Chem. Phys. **96**, 6796 (1992).

[114] D. E. Woon and T. H. J. Dunning, J. Chem. Phys. **98**, 1358 (1993).

[115] T. van Mourik, A. K. Wilson, and T. H. J. Dunning, Mol. Phys. **96**, 529 (1999).

[116] A. K. Wilson, D. E. Woon, K. A. Peterson, and T. H. J. Dunning, J. Chem. Phys. **110**, 7667 (1999).

[117] T. van Mourik and T. H. J. Dunning, Int. J. Quantum Chem. **76**, 205 (2000).

[118] D. E. Woon and T. H. J. Dunning, J. Chem. Phys. **100**, 2975 (1994).

[119] T. H. J. Dunning, K. A. Peterson, and A. K. Wilson, J. Chem. Phys. **114**, 9244 (2001).

[120] A. Schaefer, H. Horn, and R. Ahlrichs, J. Chem. Phys. **97**, 2571 (1992).

[121] A. Schaefer, C. Huber, and R. Ahlrichs, J. Chem. Phys. **100**, 5829 (1994).

[122] F. Weigend, Phys. Chem. Chem. Phys. **8**, 1057 (2006).

[123] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **78**, 1396 (1997).

[124] D. Feller, J. Chem. Phys. **96**, 6104 (1992).

[125] D. Feller, J. Chem. Phys. **98**, 7059 (1993).

[126] K. A. Peterson, D. E. Woon, and T. H. D. Jr., J. Chem. Phys. **100**, 7410 (1994).

[127] F. Jensen, Theor. Chem. Acc. **113**, 267 (2005).

[128] A. Karton and J. M. L. Martin, Theor. Chem. Acc. **115**, 330 (2006).

[129] W. Kutzelnigg and J. D. M. III, J. Chem. Phys. **96**, 4484 (1992).

[130] W. Kutzelnigg and J. D. M. III, J. Chem. Phys. **97**, 8821 (1992).

[131] C. Schwartz, Phys. Rev. **126**, 1015 (1962).

[132] J. M. L. Martin and P. R. Taylor, J. Chem. Phys. **106**, 8620 (1997).

[133] J. M. L. Martin, Chem. Phys. Lett. **259**, 679 (1996).

[134] T. Helgaker, W. Klopper, H. Koch, and J. Noga, J. Chem. Phys. **106**, 9639 (1997).

[135] E. F. Valeev, W. D. Allen, R. Hernandez, C. D. Sherrill, and H. F. S. III, J. Chem. Phys. **118**, 8594 (2003).

[136] D. W. Schwenke, J. Chem. Phys. **122**, 014107 (2005).

[137] A. Halkier, T. Helgaker, P. Jørgensen, W. Klopper, and J. Olsen, Chem. Phys. Lett. **302**, 437 (1999).

[138] A. K. Wilson and T. H. J. Dunning, J. Chem. Phys. **106**, 8718 (1997).

[139] A. T. ans L. Romaner, O. T. Hofmann, E. Zojer, C. Ambrosch-Draxl, and M. Scheffler, MRS Bull. **35**, 435 (2010).

[140] S. Grimme, A. Hansen, J. G. Brandenburg, and C. Bannwarth, Chem. Rev. **116**, 5105 (2016).

[141] A. D. Boese and N. C. Handy, J. Chem. Phys. **116**, 9559 (2002).

[142] J. D. Chai and M. Head-Gordon, J. Chem. Phys. **128**, 0841061 (2004).

[143] S. Grimme, J. Chem. Phys. **124**, 0341081 (2006).

[144] V. I. Lebedev and D. N. Laikov, Doklady Math. **59**, 477 (1999).

[145] S. Grimme, J. Comput. Chem. **27**, 1787 (2006).

[146] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, J. Chem. Phys. **132**, 154104 (2010).

[147] J.Almlof, K. Faegri, and K. Korsell, J. Comput. Chem. **3**, 385 (1982).

[148] M. Haser and R. Ahlrichs, J. Comput. Chem. **10**, 104 (1989).

[149] H. S. Yu, X. He, and D. G. Truhlar, J. Chem. Theor. Comput. **12**, 1280 (2016).

[150] G. A. Petersson, A. K. Yee, and A. Bennett, J. Chem. Phys. **83**, 5105 (1985).

[151] G. A. Petersson and M. A. Al-Laham, J. Chem. Phys. **1991**, 6081 (9).

[152] A. K. Wilson, T. van Mourik, and T. H. D. Jr., J. Mol. Struct. (Theochem) **338**, 339 (1996).

[153] C. Hättig, Phys. Chem. Chem. Phys. **7**, 59 (2005).

[154] D. E. Bernholdt and R. J. Harrison, Chem. Phys. Lett. **250**, 477 (1996).

[155] D. Quinñnero, C. Garau, A. Frontera, P. Ballester, A. Costa, and P. M. Deyè, J. Phy. Chem. A **109**, 4636 (2005).

[156] M. Goldey, A. D. Dutoi, and M. Head-Gordon, Phys. Chem. Chem. Phys. **15**, 15869 (2013).

[157] A. D. Dutoi and M. Head-Gordon, J. Phy. Chem. A **112**, 2110 (2008).

[158] P. M. W. Gill, Adv. Quantum Chem. **25**, 141 (1994).

[159] P. M. W. Gill and R. D. Adamson, Chem. Phys. Lett. **261**, 105 (1996).

[160] R. D. Adamson, J. P. Dombroski, and P. M. Gill, Chem. Phys. Lett. **254**, 329 (1996).

[161] P. M. W. Gill, Chem. Phys. Lett. **270**, 193 (1997).

[162] A. M. Lee, S. W. Taylor, J. P. Dombroski, and P. M. W. Gill, Phys. Rev. A **55**, 3233 (1997).

[163] J. P. Dombroski, S. W. Taylor, and P. M. W. Gill, J. Phys. Chem. **100**, 6272 (1996).

[164] R. D. Adamson and P. M. W. Gill, J. Mol. Struct.: THEOCHEM **45**, 398 (1997).

[165] S. M. Cybulski and M. L. Lytle, J. Chem. Phys. **127**, 141102 (2007).

[166] J. Řezáč, K. E. Riley, and P. Hobza, J. Chem. Theory Comput. **7**, 2427 (2011).

[167] H. Valdes, K. Pluhackova, M. Pitoňák, J. Rčzáč, and P. Hobza, Phys. Chem. Chem. Phys. **10**, 2747 (2008).

[168] L. M. Salonen, M. Ellermann, and F. Diederich, Angew. Chem. Int. Ed. **50**, 4808 (2011).

[169] K. E. Riley and P. Hobza, Acc. Chem. Res. **46**, 927 (2013).

[170] S. Grimme, Angew. Chem. Int. Ed. **47**, 3430 (2008).

[171] E. M. Pérez and N. Martín, Chem. Soc. Rev. **44**, 6425 (2015).

[172] H. W. Kroto, J. R. Heath, S. C. O'Brien, R. F. Curl, and R. E. Smalley, Nature **318**, 162 (1985).

[173] Y.-T. Wu and J. Siegel, Top. Curr. Chem. **349**, 63 (2014).

[174] V. M. Tsefrikas and L. T. Scott, Chem. Rev. **106**, 4868 (2006).

[175] A. Sygula, H. E. Folsom, R. Sygula, A. H. Abdourazak, Z.Marcinow, F. R. Fronczek, and P. W. Rabideau, J. Chem. Soc., Chem. Commun. **22**, 2571 (1994).

[176] D. M. Forkey, S. Attar, B. C. Noll, R. Koerner, M. M. Olmstead, and A. L. Balch, J. Am. Chem. Soc. **119**, 5766 (1997).

[177] M. A. Petrukhina, K. W. Andreini, L. Peng, and L. T. Scott, Angew. Chem. Int. Ed. **43**, 5477 (2004).

[178] A. S. Filatov, L. T. Scott, and M. A. Petrukhina, Cryst. Growth Des. **10** (2010).

[179] H. Sakurai, T. Daiko, H. Sakane, T. Amaya, and T. Hirao, J. Am. Chem. Soc. **127**, 11580 (2005).

[180] D. Canevet and N. M. E. M. Pérez, Angew. Chem. Int. Ed. **50**, 9248 (2011).

[181] T. Kawase and H. Kurata, Chem. Rev. **106**, 5250 (2006).

[182] L. N. Dawe, T. A. AlHujran, H.-A. Tran, J. I. Mercer, E. A. Jackson, L. T. Scott, and P. E. Georghiou, Chem. Commun. **48**, 5563 (2012).

[183] B. T. King, M. M. Olmstead, K. K. Baldridge, B. Kumar, A. L. Balch, and J. A. Gharamaleki, Chem. Commun. **48**, 9882 (2012).

[184] A. S. Filatov, M. V. Ferguson, S. N. Spisak, B. Li, C. F. Campana, and M. A. Petrukhina, Cryst. Growth Des. **14**, 756 (2014).

[185] W. E. Barth and R. G. Lawton, J. Am. Chem. Soc. **93**, 1730 (1971).

[186] H. Becker, G. Javahery, S. Petrie, P. C. Cheng, H. Schwarz, L. T. Scott, and D. K. Bohme, J. Am. Chem. Soc. **115**, 11636 (1993).

[187] W. Xiao, D. Passerone, P. Ruffieux, K. Aït-Mansour, O. Grönig, E. Tosatti, J. S. Siegel, and R. Fasel, J. Am. Chem. Soc. **130**, 4767 (2008).

[188] S. Mizyed, P. E. Georghiou, M. Bancu, B. Cuadra, A. K. Rai, P. Cheng, and L. T. Scott, J. Am. Chem. Soc. **123**, 12770 (2001).

[189] P. E. Georghiou, A. H. Tran, S. Mizyed, M. Bancu, and L. T. Scott, J. Org. Chem. **70**, 6158 (2005).

[190] H. Yokoi, Y. Hiraoka, S. Hiroto, D. Sakamaki, S. Seki, and H. Shinokubo, Nat. Commun. **6**, 8215 (2015).

[191] A. Sygula, F. R. Fronczek, R. Sygula, P. W. Rabideau, and M. M. Olmstead, J. Am. Chem. Soc. **129**, 3842 (2007).

[192] V. H. Le, M. Yanney, M. McGuire, A. Sygula, and E. A. Lewis, J. Phys. Chem. B **118**, 11956 (2014).

[193] M. Yanney, F. R. Fronczek, and A. Sygula, Angew. Chem. Int. Ed. **54**, 11153 (2015).

[194] P. L. A. Kuragama, F. R. Fronczek, and A. Sygula, Org. Lett. **17**, 5292 (2015).

[195] E. A. Jackson, B. D. Steinberg, M. Bancu, A. Wakamiya, and L. Scott, J. Am. Chem. Soc. **129**, 484 (2007).

[196] S. Lampart, L. M. Roch, A. K. Dutta, R. Warshamanage, A. D. Finke, A. Linden, K. K. Baldridge, and J. S. Siegel, Angew. Chem. Int. Ed. (DOI: 10.1002/anie.201608337 and 10.1002/ange.201608337).

[197] A. D. Becke, J. Chem. Phys. **107**, 8554 (1997).

[198] J. P. Perdew, J. Tao, V. N. Staroverov, and G. E. Scuseria, Phys. Rev. Lett. **91**, 1464011 (2003).

[199] J. P. Perdew, J. Tao, V. N. Staroverov, and G. E. Scuseria, J. Chem. Phys. **120**, 6898 (2004).

[200] A. D. McLean and G. S. Chandler, J. Chem. Phys. **72**, 5639 (1980).

[201] K. Raghavachari, J. S. Binkley, R. Seeger, and J. A. Pople, J. Chem. Phys. **72**, 650 (1980).

[202] A. Klamt and G. Schüürmann, J. Chem. Sco. Perkin Trans. **5**, 799 (1993).

[203] K. K. Balridge and A. Klamt, J. Chem. Phys. **106**, 6622 (1997).

[204] R. Bauernschmitt and R. Ahlrichs, Chem. Phys. Lett. **256**, 269 (1996).

[205] M. E. Casida, C. Jamorski, K. C. Casida, and D. R. Salahub, J. Chem. Phys. **108**, 4439 (1998).

[206] R. E. Stratmann, G. E. Scuseria, and M. J. Frisch, J. Chem. Phys. **109**, 8218 (1998).

[207] Y. Tawada, T. Tsuneda, S. Yanagisawa, Y. Yanai, , and K. Hirao, J. Chem. Phys. **120**, 8425 (2004).

[208] K. A. Nguyen, P. N. Day, and R. Pachter, Int. J. Quantum Chem. **110**, 2247 (2012).

[209] P. Elliott, F. Furche, and K. Burke, Rev. Comp. Chem. **26**, 91 (2009).

[210] S. Hirata and M. Head-Gordon, Chem. Phys. Lett. **314**, 291 (1999).

[211] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox, "Gaussian 09 Revision E.01," Gaussian Inc. Wallingford CT 2009.

[212] T. A. Keith and R. F. W. Bader, Chem. Phys. Lett. **194**, 1 (1992).

[213] T. A. Keith and R. F. W. Bader, Chem. Phys. Lett. **210**, 223 (1993).

[214] B. M. Bode and M. S. Gordon, J. Mol. Graphics and Modeling **16**, 133 (1999).

[215] C. F. Macrae, I. J. Bruno, J. A. Chisholm, P. R. E. P. McCabe, E. Pidcock, L. Rodriguez-Monge, R. Taylor, J. van de Streek, and P. A. Wood, J. Appl. Cryst. **41**, 466 (2008).

[216] M. A. Petrukhina, K. W. Andreini, L. Peng, and L. T. Scott, Angew. Chem. **116**, 5593 (2004).

[217] P. Job, Ann. Chim. **9**, 113 (1928).

[218] J. S. Renny, L. L. Tomasevich, E. H. Tallmadge, and D. B. Collum, Angew. Chem. **125**, 12218 (2013).

[219] J. S. Renny, L. L. Tomasevich, E. H. Tallmadge, and D. B. Collum, Angew. Chem. Int. Ed. **52**, 11998 (2013).

[220] P. A. Denis, Chem. Phys. Lett. **516**, 82 (2011).

[221] P. A. Denis, Chem. Phys. Lett. **591**, 323 (2014).

[222] D. B. Chesnut, Chem. Phys. **224**, 131 (1997).

[223] K. K. Balridge and J. S. Siegel, J. Phys. Chem. A **103**, 4038 (1999).

[224] B. D. Steinberg, E. A. Jackson, A. S. F. A. Wakamiya, M. A. Petrukhina, and L. T. Scott, J. Am. Chem. Soc. **131**, 10537 (2009).

[225] A. K. Dutta, A. Linden, L. Zoppi, K. K. Baldridge, and J. S. Siegel, Angew. Chem. **127**, 10942 (2015).

[226] A. K. Dutta, A. Linden, L. Zoppi, K. K. Baldridge, and J. S. Siegel, Angew. Chem. Int. Ed. **54**, 10792 (2015).

[227] A. Steinauer, A. M. Butterfield, A. Linden, A. Molina-Ontario, D. C. Buck, R. W. Cotta, L. Echegoyen, K. K. Baldridge, and J. S. Siegel, J. Braz. Chem. Soc. , doi: 10.5935/0103 (2016).

[228] J. Plötner, D. J. Tozer, and A. Dreuw, Journal of Chemical Theory and Computation **6**, 2315 (2010), http://pubs.acs.org/doi/pdf/10.1021/ct1001973 .

[229] M. Caricato, G. W. Trucks, M. J. Frisch, and K. B. Wiberg, Journal of Chemical Theory and Computation **6**, 370 (2010), http://pubs.acs.org/doi/pdf/10.1021/ct9005129 .

[230] D. Jacquemin, V. Wathelet, E. A. Perpete, and C. Adamo, J. Chem. Theory Comput. **5**, 2420 (2009).

[231] M. J. G. Peach, P. Benfield, T. Helgaker, and D. J. Tozer, J. Chem. Phys. **128**, 044118 (2008).

[232] M. J. G. Peach, M. J. Williamson, and D. J. Tozer, Journal of Chemical Theory and Computation **7**, 3578 (2011), http://pubs.acs.org/doi/pdf/10.1021/ct200651r .

[233] M. Caricato, G. W. Trucks, M. J. Frisch, and K. B. Wiberg, Journal of Chemical Theory and Computation **7**, 456 (2011), http://pubs.acs.org/doi/pdf/10.1021/ct100662n .

[234] N. Harada, S. L. Chen, and K. Nakanishi, J. Am. Chem. Soc. **97**, 5345 (1975).

[235] T. B. Aleksiev, *Progettazione e realizzazione di un sistema HPC altamente disponibile per high performance computing*, Master's thesis, Università degli Studi di Udine, and University of Zürich (2016).

[236] `http://www.drbd.org/en/`.

[237] `http://clusterlabs.org`.

[238] `https://www.asg.com/Journey/Atempo/Home.aspx`.

[239] `https://sourceware.org/lvm2/`.

[240] `https://www.gluster.org`.

[241] M. A. Jette, A. B. Yoo, and M. Grondona, *SLURM: Simple Linux Utility for Resource Management*, Tech. Rep. (Lawrence Livermore National Laboratory, 2002).

[242] `http://slurm.schedmd.com/cgroups.html`.

[243] `http://fai-project.org`.

[244] `https://www.ansible.com`.

[245] `http://www.linux-nis.org` (2016).

[246] `https://mathias-kettner.de/check_mk.html`.